

**PCT**WORLD INTELLECTUAL PROPERTY ORGANIZATION  
International Bureau

## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

<b>(51) International Patent Classification <sup>6</sup> :</b> <b>C12Q 1/68, C12N 9/08, A61K 51/00, C07K 1/00</b>	<b>A1</b>	<b>(11) International Publication Number:</b> <b>WO 95/20678</b> <b>(43) International Publication Date:</b> 3 August 1995 (03.08.95)
<b>(21) International Application Number:</b> PCT/US95/01035 <b>(22) International Filing Date:</b> 25 January 1995 (25.01.95)  <b>(30) Priority Data:</b> 08/187,757 27 January 1994 (27.01.94) US 08/210,143 16 March 1994 (16.03.94) US 08/294,312 23 August 1994 (23.08.94) US  <b>(71) Applicant:</b> HUMAN GENOME SCIENCES, INC. [US/US]; 9410 Key West Avenue, Rockville, MD 20850-338 (US).  <b>(72) Inventors:</b> HASELTINE, William, A.; 3035 P Street, N.W., Washington, DC 20007 (US). RUBEN, Steven, M.; 18528 Heritage Hills Drive, Olney, MD 20832 (US). WEI, Ying- Fei; 13524 Straw Bale Lane, Darnestown, MD 20878 (US). ADAMS, Mark, D.; 15205 Dufief Drive, North Potomac, MD 20878 (US). FLEISCHMANN, Robert, D.; 470 Tschiffely Square Road, Gaithersburg, MD 20878 (US). FRASER, Claire, M.; 11915 Glen Mill Road, Potomac, MD 20854 (US). FULDNER, Rebecca, A.; Box 306, 18040 Barnesville Road, Barnesville, MD 20838 (US). KIRKNESS, Ewen, F.; 2519 Little Vista Terrace, Olney, MD 20832 (US). ROSEN, Craig, A.; 22400 Rolling Hill Road, Laytonsville, MD 20882 (US).	<b>(74) Agents:</b> OLSTEIN, Elliot, M. et al.; Carella, Byrne, Bain, Gilfillan, Cecchi, Stewart & Olstein, 6 Becker Farm Road, Roseland, NJ 07068 (US).  <b>(81) Designated States:</b> AU, BB, BG, BR, CA, CN, CZ, FI, HU, JP, KG, KR, LK, MW, MX, NO, NZ, PL, RO, RU, SI, SK, UA, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).  <b>Published</b> <i>With international search report.</i> <i>Before the expiration of the time limit for amending the</i> <i>claims and to be republished in the event of the receipt of</i> <i>amendments.</i>	
<b>(54) Title:</b> HUMAN DNA MISMATCH REPAIR PROTEINS  <b>(57) Abstract</b>  The present invention discloses three human DNA repair proteins and DNA (RNA) encoding such proteins and a procedure for producing such proteins by recombinant techniques. One of the human DNA repair proteins, hMLH1, has been mapped to chromosome 3 while hMLH2 has been mapped to chromosome 2 and hMLH3 has been mapped to chromosome 7. The invention provides methods to diagnose alterations in the hMLH1, hMLH2 and hMLH3 genes.		

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgyzstan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo	KR	Republic of Korea	SE	Sweden
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LU	Luxembourg	TD	Chad
CS	Czechoslovakia	LV	Latvia	TG	Togo
CZ	Czech Republic	MC	Monaco	TJ	Tajikistan
DE	Germany	MD	Republic of Moldova	TT	Trinidad and Tobago
DK	Denmark	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	US	United States of America
FI	Finland	MN	Mongolia	UZ	Uzbekistan
FR	France			VN	Viet Nam
GA	Gabon				

## HUMAN DNA MISMATCH REPAIR PROTEINS

This invention relates to newly identified polynucleotides, polypeptides encoded by such polynucleotides, the use of such polynucleotides and polypeptides, as well as the production of such polynucleotides and polypeptides. More particularly, the polypeptides of the present invention are human homologs of the prokaryotic *mutL4* gene and are hereinafter referred to as hMLH1, hMLH2 and hMLH3.

In both prokaryotes and eukaryotes, the DNA mismatch repair gene plays a prominent role in the correction of errors made during DNA replication and genetic recombination. The *E.coli* methyl-directed DNA mismatch repair system is the best understood DNA mismatch repair system to date. In *E.coli*, this repair pathway involves the products of the mutator genes *mutS*, *mutL*, *mutH*, and *uvrD*. Mutants of any one of these genes will reveal a mutator phenotype. *MutS* is a DNA mismatch-binding protein which initiates this repair process, *uvrD* is a DNA helicase and *MutH* is a latent

endonuclease that incises at the unmethylated strands of a hemi-methylated GATC sequence. *MutL* protein is believed to recognize and bind to the mismatch-DNA-MutS-MutH complex to enhance the endonuclease activity of *MutH* protein. After the unmethylated DNA strand is cut by the *MutH*, single-stranded DNA-binding protein, DNA polymerase III, exonuclease I and DNA ligase are required to complete this repair process (Modrich P., Annu. Rev. Genetics, 25:229-53 (1991)).

Elements of the *E.coli* *MutLHS* system appears to be conserved during evolution in prokaryotes and eukaryotes. Genetic study analysis suggests that *Saccharomyces cerevisiae* has a mismatch repair system similar to the bacterial *MutLHS* system. In *S. cerevisiae*, at least two *MutL* homologs, *PMS1* and *MLH1*, have been reported. Mutation of either one of them leads to a mitotic mutator phenotype (Prolla et al., Mol. Cell. Biol. 14:407-415 (1994)). At least three *MutS* homologs have been found in *S.cerevisiae*, namely *MSH1*, *MSH2*, and *MSH3*. Disruption of the *MSH2* gene affects nuclear mutation rates. Mutants in *S. cerevisiae*, *MSH2*, *PMS1*, and *MLH1* have been found to exhibit increased rates of expansion and contraction of dinucleotide repeat sequences (Strand et al., Nature, 365:274-276 (1993)).

It has been reported that a number of human tumors such as lung cancer, prostate cancer, ovarian cancer, breast cancer, colon cancer and stomach cancer show instability of repeated DNA sequences (Han et al., Cancer, 53:5087-5089 (1993); Thibodeau et al., Science 260:816-819 (1993); Risinger et al., Cancer 53:5100-5103 (1993)). This phenomenon suggests that lack of the DNA mismatch repair is probably the cause of these tumors.

Little was known about the DNA mismatch repair system in humans until recently, the human homolog of the *MutS* gene was cloned and found to be responsible for hereditary nonpolyposis colon cancer (HNPCC), (Fishel et al., Cell, 75:1027-1038 (1993) and Leach et al., Cell, 75:1215-1225

(1993)). HNPCC was first linked to a locus at chromosome 2p16 which causes dinucleotide instability. It was then demonstrated that a DNA mismatch repair protein (MutS) homolog was located at this locus, and that C-->T transitional mutations at several conserved regions were specifically observed in HNPCC patients. Hereditary nonpolyposis colorectal cancer is one of the most common hereditary diseases of man, affecting as many as one in two hundred individuals in the western world.

It has been demonstrated that hereditary colon cancer can result from mutations in several loci. Familial adenomatous polyposis coli (APC), linked to a gene on chromosome 5, is responsible for a small minority of hereditary colon cancer. Hereditary colon cancer is also associated with Gardner's syndrome, Turcot's syndrome, Peutz-Jaegers syndrome and juvenile polyposis coli. In addition, hereditary nonpolyposis colon cancer may be involved in 5% of all human colon cancer. All of the different types of familial colon cancer have been shown to be transmitted by a dominant autosomal mode of inheritance.

In addition to localization of HNPCC, to the short arm of chromosome 2, a second locus has been linked to a predisposition to HNPCC (Lindholm, et al., Nature Genetics, 5:279-282 (1993)). A strong linkage was demonstrated between a polymorphic marker on the short arm of chromosome 3 and the disease locus.

This finding suggests that mutations on various DNA mismatch repair proteins probably play crucial roles in the development of human hereditary diseases and cancers.

HNPCC is characterized clinically by an apparent autosomal dominantly inherited predisposition to cancer of the colon, endometrium and other organs. (Lynch, H.T. et al., Gastroenterology, 104:1535-1549 (1993)). The identification of markers at 2p16 and 3p21-22 which were linked to disease in selected HNPCC kindred unequivocally

established its mendelian nature (Peltomaki, P. et al., Science, 260:810-812 (1993)). Tumors from HNPCC patients are characterized by widespread alterations of simple repeated sequences (microsatellites) (Aaltonen, L.A., et al., Science, 260:812-816 (1993)). This type of genetic instability was originally observed in a subset (12 to 18% of sporadic colorectal cancers (Id.)). Studies in bacteria and yeast indicated that a defect in DNA mismatch repair genes can result in a similar instability of microsatellites (Levinson, G. and Gutman, G.A., Nuc. Acids Res., 15:5325-5338 (1987)), and it was hypothesized that deficiency in mismatched repair was responsible for HNPCC (Strand, M. et al., Nature, 365:274-276 (1993)). Analysis of extracts from HNPCC tumor cell lines showed mismatch repair was indeed deficient, adding definitive support to this conjecture (Parsons, R.P., et al., Cell, 75:1227-1236 (1993)). As not all HNPCC kindred can be linked to the same loci, and as at least three genes can produce a similar phenotype in yeast, it seems likely that other mismatch repair genes could play a role in some cases of HNPCC.

hMLH1 is most homologous to the yeast mutL-homolog yMLH1 while hMLH2 and hMLH3 have greater homology to the yeast mutL-homolog yPMS1 (hMLH2 and hMLH3 due to their homology to yeast PMS1 gene are sometimes referred to in the literature as hPMS1 and hPMS2). In addition to hMLH1, both the hMLH2 gene on chromosome 2q32 and the hMLH3 gene, on chromosome 7p22, were found to be mutated in the germ line of HNPCC patients. This doubles the number of genes implicated in HNPCC and may help explain the relatively high incidence of this disease.

In accordance with one aspect of the present invention, there are provided novel putative mature polypeptides which are hMLH1, hMLH2 and hMLH3, as well as biologically active and diagnostically or therapeutically useful fragments,

analogs and derivatives thereof. The polypeptides of the present invention are of human origin.

In accordance with another aspect of the present invention, there are provided isolated nucleic acid molecules encoding such polypeptides, including mRNAs, DNAs, cDNAs, genomic DNA as well as biologically active and diagnostically or therapeutically useful fragments, analogs and derivatives thereof.

In accordance with still another aspect of the present invention there are provided nucleic acid probes comprising nucleic acid molecules of sufficient length to specifically hybridize to hMLH1, hMLH2 and hMLH3 sequences.

In accordance with yet a further aspect of the present invention, there is provided a process for producing such polypeptides by recombinant techniques which comprises culturing recombinant prokaryotic and/or eukaryotic host cells, containing an hMLH1, hMLH2 or hMLH3 nucleic acid sequence, under conditions promoting expression of said protein and subsequent recovery of said proteins.

In accordance with yet a further aspect of the present invention, there is provided a process for utilizing such polypeptide, or polynucleotide encoding such polypeptide, for therapeutic purposes, for example, for the treatment of cancers.

In accordance with another aspect of the present invention there is provided a method of diagnosing a disease or a susceptibility to a disease related to a mutation in the hMLH1, hMLH2 or hMLH3 nucleic acid sequences and the proteins encoded by such nucleic acid sequences.

In accordance with yet a further aspect of the present invention, there is provided a process for utilizing such polypeptides, or polynucleotides encoding such polypeptides, for in vitro purposes related to scientific research, synthesis of DNA and manufacture of DNA vectors.

These and other aspects of the present invention should be apparent to those skilled in the art from the teachings herein.

The following drawings are illustrative of embodiments of the invention and are not meant to limit the scope of the invention as encompassed by the claims.

Figure 1 illustrates the cDNA sequence and corresponding deduced amino acid sequence for the human DNA repair protein hMLH1. The amino acids are represented by their standard one-letter abbreviations. Sequencing was performed using a 373 Automated DNA sequencer (Applied Biosystems, Inc.). Sequencing accuracy is predicted to be greater than 97% accurate.

Figure 2 illustrates the cDNA sequence and corresponding deduced amino acid sequence of hMLH2. The amino acids are represented by their standard one-letter abbreviations.

Figure 3 illustrates the cDNA sequence and corresponding deduced amino acid sequence of hMLH3. The amino acids are represented by their standard one-letter abbreviations.

Figure 4. Alignment of the predicted amino acid sequences of *S. cerevisiae* PMS1 (yPMS1), with the hMLH2 and hMLH3 amino acid sequences using MACAW (version 1.0) program. Amino acid in conserved blocks are capitalized and shaded on the mean of their pair-wise scores.

Figure 5. Mutational analysis of hMLH2. (A) IVSP analysis and mapping of the transcriptional stop mutation in HNPCC patient CW. Translation of codons 1 to 369 (lane 1), codons 1 to 290 (lane 2), and codons 1 to 214 (lane 3). CW is translated from the cDNA of patient CW, while NOR was translated from the cDNA of a normal individual. The arrowheads indicate the truncated polypeptide due to the potential stop mutation. The arrows indicate molecular weight markers in kilodaltons. (B) Sequence analysis of CW indicates a C to T transition at codon 233 (indicated by the arrow). Lanes 1 and 3 are sequence derived from control



patients; lane 2 is sequence derived from genomic DNA of CW. The ddA mixes from each sequencing mix were loaded in adjacent lanes to facilitate comparison as were those for ddC, ddD, and ddT mixes.

Figure 6. Mutational analysis of hMLH3. (A) IVSP analysis of hMLH3 from patient GC. Lane GC is from fibroblasts of individual GC; lane GCx is from the tumor of patient GC; lanes NOR1 and 2 are from normal control individuals. FL indicates full-length protein, and the arrowheads indicate the germ line truncated polypeptide. The arrows indicate molecular weight markers in kilodaltons (B) PCR analysis of DNA from a patient GC shows that the lesion is present in both hMLH3 alleles in tumor cells. Amplification was done using primers that amplify 5', 3', or within (MID) the region deleted in the cDNA. Lane 1, DNA derived from fibroblasts of patient GC; lane 2, DNA derived from tumor of patient GC; lane 3, DNA derived from a normal control patient; lane 4, reactions without DNA template. Arrows indicate molecular weight in base pairs.

In accordance with an aspect of the present invention, there are provided isolated nucleic acids (polynucleotides) which encode for the mature polypeptides having the deduced amino acid sequence of Figures 1, 2 and 3 (SEQ ID No. 2, 4 and 6) or for the mature polypeptides encoded by the cDNA of the clone deposited as ATCC Deposit No. 75649, 75651, 75650, deposited on January 25, 1994.

ATCC Deposit No. 75649 is a cDNA clone which contains the full length sequence encoding the human DNA repair protein referred to herein as hMLH1; ATCC Deposit No. 75651 is a cDNA clone containing the full length cDNA sequence encoding the human DNA repair protein referred to herein as hMLH2; ATCC Deposit No. 75650 is a cDNA clone containing the full length DNA sequence referred to herein as hMLH3.

Polynucleotides encoding the polypeptides of the present invention may be obtained from one or more libraries prepared

from heart, lung, prostate, spleen, liver, gallbladder, fetal brain and testes tissues. The polynucleotides of hMLH1 were discovered from a human gallbladder cDNA library. In addition, six cDNA clones which are identical to the hMLH1 at the N-terminal ends were obtained from human cerebellum, eight-week embryo, fetal heart, HSC172 cells and Jurket cell cDNA libraries. The hMLH1 gene contains an open reading frame of 756 amino acids encoding for an 85kD protein which exhibits homology to the bacterial and yeast *mutL* proteins. However, the 5' non-translated region was obtained from the cDNA clone obtained from the fetal heart for the purpose of extending the non-translated region to design the oligonucleotides.

The hMLH2 gene was derived from a human T-cell lymphoma cDNA library. The hMLH2 cDNA clone identified an open reading frame of 2,796 base pairs flanked on both sides by in-frame termination codons. It is structurally related to the yeast PMS1 family. It contains an open reading frame encoding a protein of 934 amino acid residues. The protein exhibits the highest degree of homology to yeast PMS1 with 27% identity and 82 % similarity over the entire protein.

A second region of significant homology among the three PMS related proteins is in the carboxyl terminus, between codons 800 to 900. This region shares a 22% and 47% homology between yeast PMS1 protein and hMLH2 and hMLH3 proteins, respectively, while very little homology of this region was observed between these proteins, and the other yeast *mutL* homolog, yMLH1.

The hMLH3 gene was derived from a human endometrial tumor cDNA library. The hMLH3 clone identified a 2,586 base pair open reading frame. It is structurally related to the yPMS2 protein family. It contains an open reading frame encoding a protein of 862 amino acid residues. The protein exhibits the highest degree of homology to yPMS2 with 32%

identity and 66% similarity over the entire amino acid sequence.

It is significant with respect to a putative identification of hMLH1, hMLH2 and hMLH3 that the GFRGEAL domain which is conserved in *mutL* homologs derived from *E. coli* is conserved in the amino acid sequences of , hMLH1, hMLH2 and hMLH3.

The polynucleotides of the present invention may be in the form of RNA or in the form of DNA, which DNA includes cDNA, genomic DNA, and synthetic DNA. The DNA may be double-stranded or single-stranded, and if single stranded may be the coding strand or non-coding (anti-sense) strand. The coding sequence which encodes the mature polypeptide may be identical to the coding sequence shown in Figures 1, 2 and 3 (SEQ ID No. 1) or that of the deposited clone or may be a different coding sequence which coding sequence, as a result of the redundancy or degeneracy of the genetic code, encodes the same mature polypeptides as the DNA of Figures 1, 2 and 3 (SEQ ID No. 2, 4 and 6) or the deposited cDNA(s).

The polynucleotides which encode for the mature polypeptides of Figures 1, 2 and 3 (SEQ ID No. 2, 4 and 6) or for the mature polypeptides encoded by the deposited cDNAs may include: only the coding sequence for the mature polypeptide; the coding sequence for the mature polypeptide (and optionally additional coding sequence) and non-coding sequence, such as introns or non-coding sequence 5' and/or 3' of the coding sequence for the mature polypeptide.

Thus, the term "polynucleotide encoding a polypeptide" encompasses a polynucleotide which includes only coding sequence for the polypeptide as well as a polynucleotide which includes additional coding and/or non-coding sequence.

The present invention further relates to variants of the hereinabove described polynucleotides which encode for fragments, analogs and derivatives of the polypeptides having the deduced amino acid sequences of Figures 1, 2 and 3 (SEQ

ID No. 2, 4 and 6) or the polypeptides encoded by the cDNA of the deposited clones. The variants of the polynucleotides may be a naturally occurring allelic variant of the polynucleotides or a non-naturally occurring variant of the polynucleotides.

Thus, the present invention includes polynucleotides encoding the same mature polypeptides as shown in Figures 1, 2 and 3 (SEQ ID No. 2, 4 and 6) or the same mature polypeptides encoded by the cDNA of the deposited clones as well as variants of such polynucleotides which variants encode for a fragment, derivative or analog of the polypeptides of Figures 1, 2 and 3 (SEQ ID No. 2, 4 and 6) or the polypeptides encoded by the cDNA of the deposited clones. Such nucleotide variants include deletion variants, substitution variants and addition or insertion variants.

As hereinabove indicated, the polynucleotides may have a coding sequence which is a naturally occurring allelic variant of the coding sequence shown in Figures 1, 2 and 3 (SEQ ID No. 1, 3 and 5) or of the coding sequence of the deposited clones. As known in the art, an allelic variant is an alternate form of a polynucleotide sequence which may have a substitution, deletion or addition of one or more nucleotides, which does not substantially alter the function of the encoded polypeptide.

The polynucleotides of the present invention may also have the coding sequence fused in frame to a marker sequence which allows for purification of the polypeptides of the present invention. The marker sequence may be, for example, a hexa-histidine tag supplied by a pQE-9 vector to provide for purification of the mature polypeptides fused to the marker in the case of a bacterial host, or, for example, the marker sequence may be a hemagglutinin (HA) tag when a mammalian host, e.g. COS-7 cells, is used. The HA tag corresponds to an epitope derived from the influenza

hemagglutinin protein (Wilson, I., et al., Cell, 37:767 (1984)).

The present invention further relates to polynucleotides which hybridize to the hereinabove-described sequences if there is at least 50% and preferably 70% identity between the sequences. The present invention particularly relates to polynucleotides which hybridize under stringent conditions to the hereinabove-described polynucleotides. As herein used, the term "stringent conditions" means hybridization will occur only if there is at least 95% and preferably at least 97% identity between the sequences. The polynucleotides which hybridize to the hereinabove described polynucleotides in a preferred embodiment encode polypeptides which retain substantially the same biological function or activity as the mature polypeptides encoded by the cDNA of Figures 1, 2 and 3 (SEQ ID No. 1, 3 and 5) or the deposited cDNA(s).

The deposit(s) referred to herein will be maintained under the terms of the Budapest Treaty on the International Recognition of the Deposit of Micro-organisms for purposes of Patent Procedure. These deposits are provided merely as convenience to those of skill in the art and are not an admission that a deposit is required under 35 U.S.C. §112. The sequence of the polynucleotides contained in the deposited materials, as well as the amino acid sequence of the polypeptides encoded thereby, are incorporated herein by reference and are controlling in the event of any conflict with any description of sequences herein. A license may be required to make, use or sell the deposited materials, and no such license is hereby granted.

The present invention further relates to polypeptides which have the deduced amino acid sequence of Figures 1, 2 and 3 (SEQ ID No. 2, 4 and 6) or which have the amino acid sequence encoded by the deposited cDNA(s), as well as fragments, analogs and derivatives of such polypeptides.

The terms "fragment," "derivative" and "analog" when referring to the polypeptides of Figures 1, 2 and 3 (SEQ ID No. 2, 4 and 6) or that encoded by the deposited cDNA(s), means polypeptides which retain essentially the same biological function or activity as such polypeptides. Thus, an analog includes a proprotein which can be activated by cleavage of the proprotein portion to produce an active mature polypeptide.

The polypeptides of the present invention may be a recombinant polypeptide, a natural polypeptide or a synthetic polypeptide, preferably a recombinant polypeptide.

The fragment, derivative or analog of the polypeptides of Figures 1, 2 and 3 (SEQ ID No. 2, 4 and 6) or that encoded by the deposited cDNAs may be (i) one in which one or more of the amino acid residues are substituted with a conserved or non-conserved amino acid residue (preferably a conserved amino acid residue) and such substituted amino acid residue may or may not be one encoded by the genetic code, or (ii) one in which one or more of the amino acid residues includes a substituent group, or (iii) one in which the mature polypeptide is fused with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol). Such fragments, derivatives and analogs are deemed to be within the scope of those skilled in the art from the teachings herein.

The polypeptides and polynucleotides of the present invention are preferably provided in an isolated form, and preferably are purified to homogeneity.

The term "isolated" means that the material is removed from its original environment (e.g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or polypeptide, separated from some or all of the co-existing materials in the natural system, is isolated. Such

polynucleotides could be part of a vector and/or such polynucleotides or polypeptides could be part of a composition, and still be isolated in that such vector or composition is not part of its natural environment.

The present invention also relates to vectors which include polynucleotides of the present invention, host cells which are genetically engineered with vectors of the invention and the production of polypeptides of the invention by recombinant techniques.

Host cells are genetically engineered (transduced or transformed or transfected) with the vectors of this invention which may be, for example, a cloning vector or an expression vector. The vector may be, for example, in the form of a plasmid, a viral particle, a phage, etc. The engineered host cells can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants or amplifying the hMLH1, hMLH2 and hMLH3 genes. The culture conditions, such as temperature, pH and the like, are those previously used with the host cell selected for expression, and will be apparent to the ordinarily skilled artisan.

The polynucleotides of the present invention may be employed for producing polypeptides by recombinant techniques. Thus, for example, the polynucleotide may be included in any one of a variety of expression vectors for expressing a polypeptide. Such vectors include chromosomal, nonchromosomal and synthetic DNA sequences, e.g., derivatives of SV40; bacterial plasmids; phage DNA; baculovirus; yeast plasmids; vectors derived from combinations of plasmids and phage DNA, viral DNA such as vaccinia, adenovirus, fowl pox virus, and pseudorabies. However, any other vector may be used as long as it is replicable and viable in the host.

The appropriate DNA sequence may be inserted into the vector by a variety of procedures. In general, the DNA

sequence is inserted into an appropriate restriction endonuclease site(s) by procedures known in the art. Such procedures and others are deemed to be within the scope of those skilled in the art.

The DNA sequence in the expression vector is operatively linked to an appropriate expression control sequence(s) (promoter) to direct mRNA synthesis. As representative examples of such promoters, there may be mentioned: LTR or SV40 promoter, the E. coli lac or trp, the phage lambda P<sub>i</sub> promoter and other promoters known to control expression of genes in prokaryotic or eukaryotic cells or their viruses. The expression vector also contains a ribosome binding site for translation initiation and a transcription terminator. The vector may also include appropriate sequences for amplifying expression.

In addition, the expression vectors preferably contain one or more selectable marker genes to provide a phenotypic trait for selection of transformed host cells such as dihydrofolate reductase or neomycin resistance for eukaryotic cell culture, or such as tetracycline or ampicillin resistance in E. coli.

The vector containing the appropriate DNA sequence as hereinabove described, as well as an appropriate promoter or control sequence, may be employed to transform an appropriate host to permit the host to express the proteins.

As representative examples of appropriate hosts, there may be mentioned: bacterial cells, such as E. coli, Streptomyces, Salmonella typhimurium; fungal cells, such as yeast; insect cells such as Drosophila S2 and Spodoptera Sf9; animal cells such as CHO, COS or Bowes melanoma; adenoviruses; plant cells, etc. The selection of an appropriate host is deemed to be within the scope of those skilled in the art from the teachings herein.

More particularly, the present invention also includes recombinant constructs comprising one or more of the



sequences as broadly described above. The constructs comprise a vector, such as a plasmid or viral vector, into which a sequence of the invention has been inserted, in a forward or reverse orientation. In a preferred aspect of this embodiment, the construct further comprises regulatory sequences, including, for example, a promoter, operably linked to the sequence. Large numbers of suitable vectors and promoters are known to those of skill in the art, and are commercially available. The following vectors are provided by way of example. Bacterial: pQE70, pQE60, pQE-9 (Qiagen, Inc.), pbs, pD10, phagescript, psiX174, pbluescript SK, pbsks, pNH8A, pNH16a, pNH18A, pNH46A (Stratagene); ptrc99a, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia). Eukaryotic: pWLNEO, pSV2CAT, pOG44, pXT1, pSG (Stratagene) pSVK3, pBPV, pMSG, pSVL (Pharmacia). However, any other plasmid or vector may be used as long as they are replicable and viable in the host.

Promoter regions can be selected from any desired gene using CAT (chloramphenicol transferase) vectors or other vectors with selectable markers. Two appropriate vectors are pKK232-8 and pCM7. Particular named bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda P<sub>R</sub>, P<sub>L</sub> and TRP. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Selection of the appropriate vector and promoter is well within the level of ordinary skill in the art.

In a further embodiment, the present invention relates to host cells containing the above-described constructs. The host cell can be a higher eukaryotic cell, such as a mammalian cell, or a lower eukaryotic cell, such as a yeast cell, or the host cell can be a prokaryotic cell, such as a bacterial cell. Introduction of the construct into the host cell can be effected by calcium phosphate transfection, DEAE-Dextran mediated transfection, or electroporation (Davis, L.,

Dibner, M., Battey, I., Basic Methods in Molecular Biology, (1986)).

The constructs in host cells can be used in a conventional manner to produce the gene product encoded by the recombinant sequence. Alternatively, the polypeptides of the invention can be synthetically produced by conventional peptide synthesizers.

Mature proteins can be expressed in mammalian cells, yeast, bacteria, or other cells under the control of appropriate promoters. Cell-free translation systems can also be employed to produce such proteins using RNAs derived from the DNA constructs of the present invention. Appropriate cloning and expression vectors for use with prokaryotic and eukaryotic hosts are described by Sambrook, et al., Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor, N.Y., (1989), the disclosure of which is hereby incorporated by reference.

Transcription of the DNA encoding the polypeptides of the present invention by higher eukaryotes is increased by inserting an enhancer sequence into the vector. Enhancers are cis-acting elements of DNA, usually about from 10 to 300 bp that act on a promoter to increase its transcription. Examples including the SV40 enhancer on the late side of the replication origin bp 100 to 270, a cytomegalovirus early promoter enhancer, the polyoma enhancer on the late side of the replication origin, and adenovirus enhancers.

Generally, recombinant expression vectors will include origins of replication and selectable markers permitting transformation of the host cell, e.g., the ampicillin resistance gene of *E. coli* and *S. cerevisiae* TRP1 gene, and a promoter derived from a highly-expressed gene to direct transcription of a downstream structural sequence. Such promoters can be derived from operons encoding glycolytic enzymes such as 3-phosphoglycerate kinase (PGK),  $\alpha$ -factor, acid phosphatase, or heat shock proteins, among others. The

heterologous structural sequence is assembled in appropriate phase with translation initiation and termination sequences. Optionally, the heterologous sequence can encode a fusion protein including an N-terminal identification peptide imparting desired characteristics, e.g., stabilization or simplified purification of expressed recombinant product.

Useful expression vectors for bacterial use are constructed by inserting a structural DNA sequence encoding a desired protein together with suitable translation initiation and termination signals in operable reading phase with a functional promoter. The vector will comprise one or more phenotypic selectable markers and an origin of replication to ensure maintenance of the vector and to, if desirable, provide amplification within the host. Suitable prokaryotic hosts for transformation include E. coli, Bacillus subtilis, Salmonella typhimurium and various species within the genera Pseudomonas, Streptomyces, and Staphylococcus, although others may also be employed as a matter of choice.

As a representative but nonlimiting example, useful expression vectors for bacterial use can comprise a selectable marker and bacterial origin of replication derived from commercially available plasmids comprising genetic elements of the well known cloning vector pBR322 (ATCC 37017). Such commercial vectors include, for example, pKK223-3 (Pharmacia Fine Chemicals, Uppsala, Sweden) and GEM1 (Promega Biotec, Madison, WI, USA). These pBR322 "backbone" sections are combined with an appropriate promoter and the structural sequence to be expressed.

Following transformation of a suitable host strain and growth of the host strain to an appropriate cell density, the selected promoter is induced by appropriate means (e.g., temperature shift or chemical induction) and cells are cultured for an additional period.

Cells are typically harvested by centrifugation, disrupted by physical or chemical means, and the resulting crude extract retained for further purification.

Microbial cells employed in expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents, such methods are well known to those skilled in the art.

Various mammalian cell culture systems can also be employed to express recombinant protein. Examples of mammalian expression systems include the COS-7 lines of monkey kidney fibroblasts, described by Gluzman, Cell, 23:175 (1981), and other cell lines capable of expressing a compatible vector, for example, the C127, 3T3, CHO, HeLa and BHK cell lines. Mammalian expression vectors will comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences. DNA sequences derived from the SV40 splice, and polyadenylation sites may be used to provide the required nontranscribed genetic elements.

The polypeptides can be recovered and purified from recombinant cell cultures by methods including ammonium sulfate or ethanol precipitation, acid extraction, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, affinity chromatography, hydroxylapatite chromatography and lectin chromatography. Protein refolding steps can be used, as necessary, in completing configuration of the mature protein. Finally, high performance liquid chromatography (HPLC) can be employed for final purification steps.

The polypeptides of the present invention may be a naturally purified product, or a product of chemical synthetic procedures, or produced by recombinant techniques

from a prokaryotic or eukaryotic host (for example, by bacterial, yeast, higher plant, insect and mammalian cells in culture). Depending upon the host employed in a recombinant production procedure, the polypeptides of the present invention may be glycosylated or may be non-glycosylated.

In accordance with a further aspect of the invention, there is provided a process for determining susceptibility to cancer, in particular, a hereditary cancer. Thus, a mutation in a human repair protein, which is a human homolog of *mutL*, and in particular those described herein, indicates a susceptibility to cancer, and the nucleic acid sequences encoding such human homologs may be employed in an assay for ascertaining such susceptibility. Thus, for example, the assay may be employed to determine a mutation in a human DNA repair protein as herein described, such as a deletion, truncation, insertion, frame shift, etc., with such mutation being indicative of a susceptibility to cancer.

A mutation may be ascertained for example, by a DNA sequencing assay. Tissue samples, including but not limited to blood samples are obtained from a human patient. The samples are processed by methods known in the art to capture the RNA. First strand cDNA is synthesized from the RNA samples by adding an oligonucleotide primer consisting of polythymidine residues which hybridize to the polyadenosine stretch present on the mRNA's. Reverse transcriptase and deoxynucleotides are added to allow synthesis of the first strand cDNA. Primer sequences are synthesized based on the DNA sequence of the DNA repair protein of the invention. The primer sequence is generally comprised of 15 to 30 and preferably from 18 to 25 consecutive bases of the human DNA repair gene. Table 1 sets forth an illustrative example of oligonucleotide primer sequences based on *hMLH1*. The primers are used in pairs (one "sense" strand and one "anti-sense") to amplify the cDNA from the patients by the PCR method (Saiki et al., *Nature*, 324:163-166 (1986)) such that three

overlapping fragments of the patient's cDNA's for such protein are generated. Table 1 also shows a list of preferred primer sequence pairs. The overlapping fragments are then subjected to dideoxynucleotide sequencing using a set of primer sequences synthesized to correspond to the base pairs of the cDNA's at a point approximately every 200 base pairs throughout the gene.

TABLE 1Primer Sequences used to amplify gene region using PCR

<u>Name</u>	<u>Start Site and Arrangement</u>	<u>Sequence</u>
758	sense-(-41)*	GTTGAACATCTAGACGTCTC
1319	sense-8	TCGTGGCAGGGGTTATTCTG
1321	sense-619	CTACCCAATGCCTCAACCG
1322	sense-677	GAGAACTGATAGAAATTGGATG
1314	sense-1548	GGGACATGAGGTTCTCCG
1323	sense-1593	GGGCTGTGTGAATCCTCAG
773	anti-53	CGGTTCACTACTGTCTCGTC
1313	anti-971	TCCAGGATGCTCTCCTCG
1320	anti-1057	CAAGTCCTGGTAGCAAAGTC
1315	anti-1760	ATGGCAAGGTCAAAGAGCG
1316	anti-1837	CAACAATGTATTCAGXAAGTCC
1317	anti-2340	TTGATACAACACTTTGTATCG
1318	anti-2415	GGAATACTATCAGAAGGCAAG

\* Numbers corresponding to location along nucleotide sequence of Figure 1 where ATG is number 1.  
Preferred primer sequences pairs:

758, 1313  
1319, 1320  
660, 1909  
725, 1995  
1680, 2536  
1727, 2610

The nucleotide sequences shown in Table 1 represent SEQ ID No. 7 through 19, respectively.

Table 2 lists representative examples of oligonucleotide primer sequences (sense and anti-sense) which may be used, and preferably the entire set of primer sequences are used for sequencing to determine where a mutation in the patient DNA repair protein may be. The primer sequences may be from 15 to 30 bases in length and are preferably between 18 and 25 bases in length. The sequence information determined from the patient is then compared to non-mutated sequences to determine if any mutations are present.

TABLE 2Primer Sequences Used to Sequence the Amplified Fragments

<u>Name</u>	<u>Number</u>	<u>Start Site</u> <u>and Arrangement</u>	<u>Sequence</u>
5282	seq01	sense-377*	ACAGAGCAAGTTACTCAGATG
5283	seq02	sense-552	GTACACAATGCAGGCATTAG
5284	seq03	sense-904	AATGTGGATGTTAATGTGCAC
5285	seq04	sense-1096	CTGACCTCGTCTTCCTAC
5286	seq05	sense-1276	CAGCAAGATGAGGAGATGC
5287	seq06	sense-1437	GGAAATGGTGGGAAGATGATTC
5288	seq07	sense-1645	CTTCTCAACACCAAGC
5289	seq08	sense-1895	GAAATTGATGAGGAAGGGAAC
5295	seq09	sense-1921	CTTCTGATTGACAACTATGTGC
5294	seq10	sense-2202	CACAGAAGATGGAAATATCCTG
5293	seq11	sense-2370	GTGTTGGTAGCACTTAAGAC
5291	seq12	anti-525	TTTCCCATATTCTTCACTTG
5290	seq13	anti-341	GTAACATGAGCCACATGGC
5292	seq14	anti-46	CCACTGTCTCGTCCAGCCG

\* Numbers corresponding to location along nucleotide sequence of Figure 1 where ATG is number 1.

The nucleotide sequences shown in Table 2 represent SEQ ID No. 20 through 33, respectively.

In another embodiment, the primer sequences from Table



2 could be used in the PCR method to amplify a mutated region. The region could be sequenced and used as a diagnostic to predict a predisposition to such mutated genes.

Alternatively, the assay to detect mutations in the genes of the present invention may be performed by genetic testing based on DNA sequence differences achieved by detection of alteration in electrophoretic mobility of DNA fragments in gels with or without denaturing agents. Small sequence deletions and insertions can be visualized by high resolution gel electrophoresis. DNA fragments of different sequences may be distinguished on denaturing formamide gradient gels in which the mobilities of different DNA fragments are retarded in the gel at different positions according to their specific melting or partial melting temperatures (see, e.g., Myers et al., Science, 230:1242 (1985)).

Sequence changes at specific locations may also be revealed by nuclease protection assays, such as RNase and S1 protection or the chemical cleavage method (e.g., Cotton et al., PNAS, USA, 85:4397-4401 (1985)). Perfectly matched sequences can be distinguished from mismatched duplexes by RNase A digestion or by differences in melting temperatures.

Thus, the detection of a specific DNA sequence may be achieved by methods such as hybridization, RNase protection, chemical cleavage, Western Blot analysis,

direct DNA sequencing or the use of restriction enzymes, (e.g., Restriction Fragment Length Polymorphisms (RFLP)) and Southern blotting of genomic DNA.

In addition to more conventional gel-electrophoresis and DNA sequencing, mutations can also be detected by *in situ* analysis.

The polypeptides may also be employed to treat cancers or to prevent cancers, by expression of such polypeptides *in vivo*, which is often referred to as "gene therapy."

Thus, for example, cells from a patient may be engineered with a polynucleotide (DNA or RNA) encoding a polypeptide *ex vivo*, with the engineered cells then being provided to a patient to be treated with the polypeptide. Such methods are well-known in the art. For example, cells may be engineered by procedures known in the art by use of a retroviral particle containing RNA encoding a polypeptide of the present invention.

Similarly, cells may be engineered *in vivo* for expression of a polypeptide *in vivo* by, for example, procedures known in the art. As known in the art, a producer cell for producing a retroviral particle containing RNA encoding the polypeptide of the present invention may be administered to a patient for engineering cells *in vivo* and expression of the polypeptide *in vivo*. These and other methods for administering a polypeptide of the present invention by such method should be apparent to those skilled in the art from the teachings of the present

invention. For example, the expression vehicle for engineering cells may be other than a retrovirus, for example, an adenovirus which may be used to engineer cells in vivo after combination with a suitable delivery vehicle.

Each of the cDNA sequences identified herein or a portion thereof can be used in numerous ways as polynucleotide reagents. The sequences can be used as diagnostic probes for the presence of a specific mRNA in a particular cell type. In addition, these sequences can be used as diagnostic probes suitable for use in genetic linkage analysis (polymorphisms).

The sequences of the present invention are also valuable for chromosome identification. The sequence is specifically targeted to and can hybridize with a particular location on an individual human chromosome. Moreover, there is a current need for identifying particular sites on the chromosome. Few chromosome marking reagents based on actual sequence data (repeat polymorphisms) are presently available for marking chromosomal location. The mapping of DNAs to chromosomes according to the present invention is an important first step in correlating those sequences with genes associated with disease.

Briefly, sequences can be mapped to chromosomes by preparing PCR primers (preferably 15-25 bp) from the cDNA. Computer analysis of the 3' untranslated region is used to rapidly select primers that do not span more than one exon

in the genomic DNA, thus complicating the amplification process. These primers are then used for PCR screening of somatic cell hybrids containing individual human chromosomes. Only those hybrids containing the human gene corresponding to the primer will yield an amplified fragment.

PCR mapping of somatic cell hybrids is a rapid procedure for assigning a particular DNA to a particular chromosome. Using the present invention with the same oligonucleotide primers, sublocalization can be achieved with panels of fragments from specific chromosomes or pools of large genomic clones in an analogous manner. Other mapping strategies that can similarly be used to map to its chromosome include *in situ* hybridization, prescreening with labeled flow-sorted chromosomes and preselection by hybridization to construct chromosome-specific cDNA libraries.

Fluorescence *in situ* hybridization (FISH) of a cDNA clone to a metaphase chromosomal spread can be used to provide a precise chromosomal location in one step. This technique can be used with cDNA as short as 500 or 600 bases; however, clones larger than that have a higher likelihood of binding to a unique chromosomal location with sufficient signal intensity for simple detection. FISH requires use of the clones from which the express sequence tag or EST was derived, and the longer the better. For example, 2,000 bp is good, 4,000 is better, and more than

4,000 is probably not necessary to get good results a reasonable percentage of the time. For a review of this technique, see Verma et al., Human Chromosomes: a Manual of Basic Techniques, Pergamon Press, New York (1988).

Once a sequence has been mapped to a precise chromosomal location, the physical position of the sequence on the chromosome can be correlated with genetic map data. Such data are found, for example, in V. McKusick, Mendelian Inheritance in Man (available on line through Johns Hopkins University Welch Medical Library). The relationship between genes and diseases that have been mapped to the same chromosomal region are then identified through linkage analysis (coinheritance of physically adjacent genes).

Next, it is necessary to determine the differences in the cDNA or genomic sequence between affected and unaffected individuals. If a mutation is observed in some or all of the affected individuals but not in any normal individuals, then the mutation is likely to be the causative agent of the disease.

With current resolution of physical mapping and genetic mapping techniques, a cDNA precisely localized to a chromosomal region associated with the disease could be one of between 50 and 500 potential causative genes. (This assumes 1 megabase mapping resolution and one gene per 20 kb).

hMLH2 has been localized using a genomic P1 clone (1670) which contained the 5' region of the hMLH2 gene.

Detailed analysis of human metaphase chromosome spreads, counterstained to reveal banding, indicated that the hMLH2 gene was located within bands 2q32. Likewise, hMLH3 was localized using a genomic P1 clone (2053) which contained the 3' region of the hMLH3 gene. Detailed analysis of human metaphase chromosome spreads, counterstained to reveal banding, indicated that the hMLH3 gene was located within band 7p22, the most distal band on chromosome 7. Analysis with a variety of genomic clones showed that hMLH3 was a member of a subfamily of related genes, all on chromosome 7.

The polypeptides, their fragments or other derivatives, or analogs thereof, or cells expressing them can be used as an immunogen to produce antibodies thereto. These antibodies can be, for example, polyclonal or monoclonal antibodies. The present invention also includes chimeric, single chain, and humanized antibodies, as well as Fab fragments, or the product of an Fab expression library. Various procedures known in the art may be used for the production of such antibodies and fragments.

Antibodies generated against the polypeptides corresponding to a sequence of the present invention can be obtained by direct injection of the polypeptides into an animal or by administering the polypeptides to an animal, preferably a nonhuman. The antibody so obtained will then bind the polypeptides itself. In this manner, even a sequence encoding only a fragment of the polypeptides can

be used to generate antibodies binding the whole native polypeptides. Such antibodies can then be used to isolate the polypeptide from tissue expressing that polypeptide.

For preparation of monoclonal antibodies, any technique which provides antibodies produced by continuous cell line cultures can be used. Examples include the hybridoma technique (Kohler and Milstein, 1975, Nature, 256:495-497), the trioma technique, the human B-cell hybridoma technique (Kozbor et al., 1983, Immunology Today 4:72), and the EBV-hybridoma technique to produce human monoclonal antibodies (Cole, et al., 1985, in Monoclonal Antibodies and Cancer Therapy, Alan R. Liss, Inc., pp. 77-96).

Techniques described for the production of single chain antibodies (U.S. Patent 4,946,778) can be adapted to produce single chain antibodies to immunogenic polypeptide products of this invention. Also, transgenic mice may be used to express humanized antibodies to immunogenic polypeptide products of this invention.

The present invention will be further described with reference to the following examples; however, it is to be understood that the present invention is not limited to such examples. All parts or amounts, unless otherwise specified, are by weight.

In order to facilitate understanding of the following examples certain frequently occurring methods and/or terms will be described.

"Plasmids" are designated by a lower case p preceded and/or followed by capital letters and/or numbers. The starting plasmids herein are either commercially available, publicly available on an unrestricted basis, or can be constructed from available plasmids in accord with published procedures. In addition, equivalent plasmids to those described are known in the art and will be apparent to the ordinarily skilled artisan.

"Digestion" of DNA refers to catalytic cleavage of the DNA with a restriction enzyme that acts only at certain sequences in the DNA. The various restriction enzymes used herein are commercially available and their reaction conditions, cofactors and other requirements were used as would be known to the ordinarily skilled artisan. For analytical purposes, typically 1  $\mu$ g of plasmid or DNA fragment is used with about 2 units of enzyme in about 20  $\mu$ l of buffer solution. For the purpose of isolating DNA fragments for plasmid construction, typically 5 to 50  $\mu$ g of DNA are digested with 20 to 250 units of enzyme in a larger volume. Appropriate buffers and substrate amounts for particular restriction enzymes are specified by the manufacturer. Incubation times of about 1 hour at 37°C are ordinarily used, but may vary in accordance with the supplier's instructions. After digestion the reaction is electrophoresed directly on a polyacrylamide gel to isolate the desired fragment.



Size separation of the cleaved fragments is performed using 8 percent polyacrylamide gel described by Goeddel, D. et al., Nucleic Acids Res., 8:4057 (1980).

"Oligonucleotides" refers to either a single stranded polydeoxynucleotide or two complementary polydeoxynucleotide strands which may be chemically synthesized. Such synthetic oligonucleotides have no 5' phosphate and thus will not ligate to another oligonucleotide without adding a phosphate with an ATP in the presence of a kinase. A synthetic oligonucleotide will ligate to a fragment that has not been dephosphorylated.

"Ligation" refers to the process of forming phosphodiester bonds between two double stranded nucleic acid fragments (Maniatis, T., et al., Id., p. 146). Unless otherwise provided, ligation may be accomplished using known buffers and conditions with 10 units to T4 DNA ligase ("ligase") per 0.5  $\mu$ g of approximately equimolar amounts of the DNA fragments to be ligated.

Unless otherwise stated, transformation was performed as described in the method of Graham, F. and Van der Eb, A., Virology, 52:456-457 (1973).

#### Example 1

##### Bacterial Expression of hMLH1

The full length DNA sequence encoding human DNA mismatch repair protein hMLH1, ATCC # 75649, is initially amplified using PCR oligonucleotide primers corresponding to the 5' and 3' ends of the DNA sequence to synthesize

insertion fragments. The 5' oligonucleotide primer has the sequence 5' CGGGATCCATGTCGTTTCGTGGCAGGG 3' (SEQ ID No. 34), contains a BamHI restriction enzyme site followed by 18 nucleotides of hMLH1 coding sequence following the initiation codon; the 3' sequence 5' GCTCTAGATTAAACACCTCT CAAAGAC 3' (SEQ ID No. 35) contains complementary sequences to an XbaI site and is at the end of the gene. The restriction enzyme sites correspond to the restriction enzyme sites on the bacterial expression vector pQE-9. (Qiagen, Inc., Chatsworth, CA). The plasmid vector encodes antibiotic resistance (Amp<sup>r</sup>), a bacterial origin of replication (ori), an IPTG-regulatable promoter/operator (P/O), a ribosome binding site (RBS), a 6-histidine tag (6-His) and restriction enzyme cloning sites. The pQE-9 vector is digested with BamHI and XbaI and the insertion fragments are then ligated into the pQE-9 vector maintaining the reading frame initiated at the bacterial RBS. The ligation mixture is then used to transform the *E. coli* strain M15/rep4 (Qiagen, Inc.) which contains multiple copies of the plasmid pREP4, which expresses the lacI repressor and also confers kanamycin resistance (Kan<sup>r</sup>). Transformants are identified by their ability to grow on LB plates and ampicillin/kanamycin resistant colonies are selected. Plasmid DNA is isolated and confirmed by restriction analysis. Clones containing the desired constructs are grown overnight (O/N) in liquid culture in LB media supplemented with both Amp (100 ug/ml) and Kan (25

ug/ml). The O/N culture is used to inoculate a large culture at a ratio of 1:100 to 1:250. The cells are grown to an optical density 600 (O.D.<sup>600</sup>) of between 0.4 and 0.6. IPTG (Isopropyl-B-D-thiogalacto pyranoside) is then added to a final concentration of 1 mM. IPTG induces by inactivating the lacI repressor, clearing the P/O leading to increased gene expression. Cells are grown an extra 3 to 4 hours. Cells are then harvested by centrifugation (20 mins at 6000Xg). The cell pellet is solubilized in the chaotropic agent 6 Molar Guanidine HCl. After clarification, solubilized hMLH1 is purified from this solution by chromatography on a Nickel-Chelate column under conditions that allow for tight binding by proteins containing the 6-His tag (Hochuli, E. et al., Genetic Engineering, Principles & Methods, 12:87-98 (1990)). Protein renaturation out of GnHCl can be accomplished by several protocols (Jaenicke, R. and Rudolph, R., Protein Structure - A Practical Approach, IRL Press, New York (1990)). Initially, step dialysis is utilized to remove the GnHCL. Alternatively, the purified protein isolated from the Ni-chelate column can be bound to a second column over which a decreasing linear GnHCL gradient is run. The protein is allowed to renature while bound to the column and is subsequently eluted with a buffer containing 250 mM Imidazole, 150 mM NaCl, 25 mM Tris-HCl pH 7.5 and 10% Glycerol. Finally, soluble protein is dialyzed against a

storage buffer containing 5 mM Ammonium Bicarbonate. The purified protein was analyzed by SDS-PAGE.

### Example 2

#### Spontaneous Mutation Assay for Detection of the Expression of hMLH1, hMLH2 and hMLH3 and Complementation to the *E.coli mutL*

The pQE9hMLH1, pQE9hMLH2 or pQE9hMLH3/GW3733, transformants were subjected to the spontaneous mutation assay. The plasmid vector pQE9 was also transformed to AB1157 (*k-12, argE3 hisG4, LeuB6 proA2 thr-1 ara-1 rpsL31 supE44 tsx-33*) and GW3733 to use as the positive and negative control respectively.

Fifteen 2 ml cultures, inoculated with approximately 100 to 1000 *E. coli*, were grown  $2 \times 10^8$  cells per ml in LB ampicillin medium at 37°C. Ten microliters of each culture were diluted and plated on the LB ampicillin plates to measure the number of viable cells. The rest of the cells from each culture were then concentrated in saline and plated on minimal plates lacking of arginine to measure reversion of Arg<sup>+</sup>. In Table 3, the mean number of mutations per culture ( $m$ ) was calculated from the median number ( $r$ ) of mutants per distribution, according to the equation  $(r/m) - \ln(m) = 1.24$  (Lea et al., J. Genetics 49:264-285 (1949)). Mutation rates per generation were recorded as  $m/N$ , with  $N$  representing the average number of cells per culture.

TABLE 3

## Spontaneous Mutation Rates

Strain	Mutation/generation
AB1157+vector	$(5.6 \pm 0.1) \times 10^{-9a}$
GW3733+vector	$(1.1 \pm 0.2) \times 10^{-6a}$
GW3733+phMLH1	$(3.7 \pm 1.3) \times 10^{-7a}$
GW3733+phMLH2	$(3.1 \pm 0.6) \times 10^{-7b}$
GW3733+phMLH3	$(2.1 \pm 0.8) \times 10^{-7b}$

a: Average of three experiments.

b: Average of four experiments.

The functional complementation result showed that the human *mutL* can partially rescue the *E.coli* *mutL* mutator phenotype, suggesting that the human *mutL* is not only successfully expressed in a bacterial expression system, but also functions in bacteria.

Example 3Chromosomal Mapping of the hMLH1

An oligonucleotide primer set was designed according to the sequence at the 5' end of the cDNA for HMLH1. This primer set would span a 94 bp segment. This primer set was used in a polymerase chain reaction under the following set of conditions :

30 seconds, 95 degrees C

1 minute, 56 degrees C

1 minute, 70 degrees C

This cycle was repeated 32 times followed by one 5 minute cycle at 70 degrees C. Human, mouse, and hamster DNA were used as template in addition to a somatic cell hybrid panel (Bios, Inc). The reactions were analyzed on either 8% polyacrylamide gels or 3.5 % agarose gels. A 94 base pair band was observed in the human genomic DNA sample and in the somatic cell hybrid sample corresponding to chromosome 3. In addition, using various other somatic cell hybrid genomic DNA, the hMLH1 gene was localized to chromosome 3p.

#### Example 4

##### Method for Determination of mutation of hMLH1 gene in HNPCC kindred

cdNA was produced from RNA obtained from tissue samples from persons who are HNPCC kindred and the cdNA was used as a template for PCR, employing the primers 5' GCATC TAGACGTTTCCTTGGC 3' (SEQ ID No. 36) and 5' CATCCAAGCTTCTGT TCCCG 3' (SEQ ID No. 37), allowing amplification of codons 1 to 394 of Figure 1; 5' GGGGTGCAGCAGCACATCG 3' (SEQ ID No. 38) and 5' GGAGGCAGAATGTGTGAGCG 3' (SEQ ID No. 39), allowing amplification of codons 326 to 729 of Figure 1 (SEQ ID No. 2); and 5' TCCCAAAGAAGGACTTGCT 3' (SEQ ID No. 40) and 5' AGTATAAGTCTTAAGTGCTACC 3' (SEQ ID No. 41), allowing amplification of codons 602 to 756 plus 128 nt of

3'- untranslated sequences of ~~Figure 1~~ (SEQ ID No. 2). The PCR conditions for all analyses used consisted of 35 cycles at 95°C for 30 seconds, 52-58°C for 60 to 120 seconds, and 70°C for 60 to 120 seconds, in the buffer solution described in San Sidransky, D. et al., Science, 252:706 (1991). PCR products were sequenced using primers labeled at their 5' end with T4 polynucleotide kinase, employing SequiTherm Polymerase (Epicentre Technologies). The intron-exon borders of selected exons were also determined and genomic PCR products analyzed to confirm the results. PCR products harboring suspected mutations were then cloned and sequenced to validate the results of the direct sequencing. PCR products were cloned into T-tailed vectors as described in Holton, T.A. and Graham, M.W., Nucleic Acids Research, 19:1156 (1991) and sequenced with T7 polymerase (United States Biochemical). Affected individuals from seven kindreds all exhibited a heterozygous deletion of codons 578 to 632 of the hMLH1 gene. The derivation of five of these seven kindreds could be traced to a common ancestor. The genomic sequences surrounding codons 578-632 were determined by cycle-sequencing of the P1 clones (a human genomic P1 library which contains the entire hMLH1 gene (Genome Systems)) using SequiTherm Polymerase, as described by the manufacturer, with the primers were labeled with T4 polynucleotide kinase, and by sequencing PCR products of genomic DNA. The primers used to amplify the exon

containing codons 578-632 were 5' TTTATGGTTTCTCACCTGCC 3' (SEQ ID No. 42) and 5' GTTATCTGCCCCACCTCAGC 3' (SEQ ID No. 43). The PCR product included 105 bp of intron C sequence upstream of the exon and 117 bp downstream. No mutations in the PCR product were observed in the kindreds, so the deletion in the RNA was not due to a simple splice site mutation. Codons 578 to 632 were found to constitute a single exon which was deleted from the gene product in the kindreds described above. This exon contains several highly conserved amino acids.

In a second family (L7), PCR was performed using the above primers and a 4bp deletion was observed beginning at the first nucleotide (nt) of codon 727. This produced a frame shift with a new stop codon 166 nt downstream, resulting in a substitution of the carboxy-terminal 29 amino acids of hMLH1 with 53 different amino acids, some encoded by nt normally in the 3' untranslated region.

A different mutation was found in a different kindred (L2516) after PCR using the above primers, the mutation consisting of a 4bp insert between codons 755 and 756. This insertion resulted in a frame shift and extension of the ORF to include 102 nucleotides (34 amino acids) downstream of the normal termination codon. The mutations in both kindreds L7 and L2516 were therefore predicted to alter the C-terminus of hMLH1.

A possible mutation in the hMLH1 gene was determined from alterations in size of the encoded protein, where



kindreds were too few for linkage studies. The primers used for coupled transcription-translation of hMLH1 were 5' GGATCCTAATACGACTCACTATAGGGAGACCACCATGGCATCT AGACGTTTCCCTTGGC 3' (SEQ ID No. 44) and 5' CATCCAAGCTTCTGTTCCTCG 3' (SEQ ID No. 45) for codons 1 to 394 of Figure 1 and 5' GGATCCTAATACGACTCACTATAGGGAGACCACCATGGG GGTGCAGCAGCACATCG 3' (SEQ ID No. 46) and 5' GGAGGCAGAATGTG TGAGCG 3' (SEQ ID No. 47) for codons 326 to 729 of Figure 1 (SEQ ID No. 2). The resultant PCR products had signals for transcription by T7 RNA polymerase and for the initiation of translation at their 5' ends. RNA from lymphoblastoid cells of patients from 18 kindreds was used to amplify two products, extending from codon 1 to codon 394 or from codon 326 to codon 729, respectively. The PCR products were then transcribed and translated *in vitro*, making use of transcription-translation signals incorporated into the PCR primers. PCR products were used as templates in coupled transcription-translation reactions performed as described by Powell, S.M. et al., New England Journal of Medicine, 329:1982, (1993), using 40 micro Ci of <sup>35</sup>S labeled methionine. Samples were diluted in sample buffer, boiled for five minutes and analyzed by electrophoresis on sodium dodecyl sulfate-polyacrylamide gels containing a gradient of 10% to 20% acrylamide. The gels were dried and subjected to radiography. All samples exhibited a polypeptide of the expected size, but an abnormally migrating polypeptide was additionally found in one case.

The sequence of the relevant PCR product was determined and found to include a 371 bp deletion beginning at the first nucleotide (nt) of codon 347. This alteration was present in heterozygous form, and resulted in a frame shift in a new stop codon 30 nt downstream of codon 346, thus explaining the truncated polypeptide observed.

Four colorectal tumor cell lines manifesting microsatellite instability were examined. One of the four (cell line H6) showed no normal peptide in this assay and produced only a short product migrating at 27 kd. The sequence of the corresponding cDNA was determined and found to harbor a C to A transversion at codon 252, resulting in the substitution of a termination codon for serine. In accord with the translational analyses, no band at the normal C position was identified in the cDNA or genomic DNA from this tumor, indicating that it was devoid of a functional hMLH1 gene.

Table 4 sets forth the results of these sequencing assays. Deletions were found in those people who were known to have a family history of the colorectal cancer. More particularly, 9 of 10 families showed an hMLH1 mutation.

Table 4 - Summary of Mutations in *hMLH1*

<u>Sample</u>	<u>Codon</u>	<u>cDNA Nucleotide Change</u>	<u>Predicted Coding Change</u>
Kindreds F2, F3, F6, F8, F10, F11, F52	578-632	165 bp deletion	In-frame deletion
Kindred L7	727/728	4 bp deletion (TCACACATTC to TCATTCT)	Frameshift and substitution of new amino acids
Kindred L2516	755/756	4 bp insertion (GTGTTAA to GTGTTTGTTAA)	Extension of C-terminus
Kindred RA	347	371 bp deletion	Frameshift/ Truncation
H6 Colorectal Tumor	252	Transversion (TCA to TAA)	Serine to Stop

Example 5Bacterial Expression and Purification of *hMLH2*

The DNA sequence encoding *hMLH2*, ATCC #75651, is initially amplified using PCR oligonucleotide primers corresponding to the 5' and 3' ends of the DNA sequence to synthesize insertion fragments. The 5' oligonucleotide primer has the sequence 5' CGGGATCCATGAAACAATTGCCTGCGGC 3' (SEQ ID No. 48) contains a BamHI restriction enzyme site

followed by 17 nucleotides of hMLH2 following the initiation codon. The 3' sequence 5' GCTCTAGACCAGACTCAT GCTGTTTT 3' (SEQ ID No. 49) contains complementary sequences to an XbaI site and is followed by 18 nucleotides of hMLH2. The restriction enzyme sites correspond to the restriction enzyme sites on the bacterial expression vector pQE-9 (Qiagen, Inc. Chatsworth, CA). pQE-9 encodes antibiotic resistance (Amp'), a bacterial origin of replication (ori), an IPTG-regulatable promoter operator (P/O), a ribosome binding site (RBS), a 6-His tag and restriction enzyme sites. The amplified sequences and pQE-9 are then digested with BamHI and XbaI. The amplified sequences are ligated into pQE-9 and are inserted in frame with the sequence encoding for the histidine tag and the RBS. The ligation mixture is then used to transform E. coli strain M15/rep4 (Qiagen, Inc.) which contains multiple copies of the plasmid pREP4, which expresses the lacI repressor and also confers kanamycin resistance (Kan'). Transformants are identified by their ability to grow on LB plates and ampicillin/kanamycin resistant colonies are selected. Plasmid DNA is isolated and confirmed by restriction analysis. Clones containing the desired constructs are grown overnight (O/N) in liquid culture in LB media supplemented with both Amp (100 ug/ml) and Kan (25 ug/ml). The O/N culture is used to inoculate a large culture at a ratio of 1:100 to 1:250. The cells are grown to an optical density 600 (O.D.<sup>600</sup>) of between 0.4 and 0.6.

IPTG (Isopropyl-B-D-thiogalacto pyranoside) is then added to a final concentration of 1 mM. IPTG induces by inactivating the lacI repressor, clearing the P/O leading to increased gene expression. Cells are grown an extra 3 to 4 hours. Cells are then harvested by centrifugation (20 mins at 6000Xg). The cell pellet is solubilized in the chaotropic agent 6 Molar Guanidine HCl. After clarification, solubilized hMLH2 is purified from this solution by chromatography on a Nickel-Chelate column under conditions that allow for tight binding by proteins containing the 6-His tag (Hochuli, E. et al., Genetic Engineering, Principles & Methods, 12:87-98 (1990)). Protein renaturation out of GnHCl can be accomplished by several protocols (Jaenicke, R. and Rudolph, R., Protein Structure - A Practical Approach, IRL Press, New York (1990)). Initially, step dialysis is utilized to remove the GnHCl. Alternatively, the purified protein isolated from the Ni-chelate column can be bound to a second column over which a decreasing linear GnHCl gradient is run. The protein is allowed to renature while bound to the column and is subsequently eluted with a buffer containing 250 mM Imidazole, 150 mM NaCl, 25 mM Tris-HCl pH 7.5 and 10% Glycerol. Finally, soluble protein is dialyzed against a storage buffer containing 5 mM Ammonium Bicarbonate. The purified protein was analyzed by SDS-PAGE.

Example 6Bacterial Expression and Purification of hMLH3

The DNA sequence encoding hMLH3, ATCC #75650, is initially amplified using PCR oligonucleotide primers corresponding to the 5' and 3' ends of the DNA sequence to synthesize insertion fragments. The 5' oligonucleotide primer has the sequence 5' CGGGATCCATGGAGCGAGCTGAGAGC 3' (SEQ ID No. 50) contains a BamHI restriction enzyme site followed by 18 nucleotides of hMLH3 coding sequence starting from the presumed terminal amino acid of the processed protein. The 3' sequence 5' GCTCTAGAGTGAAG ACTCTGTCT 3' (SEQ ID No. 51) contains complementary sequences to an XbaI site and is followed by 18 nucleotides of hMLH3. The restriction enzyme sites correspond to the restriction enzyme sites on the bacterial expression vector pQE-9 (Qiagen, Inc. Chatsworth, CA). pQE-9 encodes antibiotic resistance (Amp'), a bacterial origin of replication (ori), an IPTG-regulatable promoter operator (P/O), a ribosome binding site (RBS), a 6-His tag and restriction enzyme sites. The amplified sequences and pQE-9 are then digested with BamHI and XbaI. The amplified sequences are ligated into pQE-9 and are inserted in frame with the sequence encoding for the histidine tag and the RBS. The ligation mixture was then used to transform E. coli strain M15/rep4 (Qiagen, Inc.) which contains multiple copies of the plasmid pREP4, which expresses the lacI repressor and also confers kanamycin resistance (Kan').

Transformants are identified by their ability to grow on LB plates and ampicillin/kanamycin resistant colonies are selected. Plasmid DNA is isolated and confirmed by restriction analysis. Clones containing the desired constructs are grown overnight (O/N) in liquid culture in LB media supplemented with both Amp (100 ug/ml) and Kan (25 ug/ml). The O/N culture is used to inoculate a large culture at a ratio of 1:100 to 1:250. The cells are grown to an optical density 600 (O.D.<sup>600</sup>) of between 0.4 and 0.6. IPTG (Isopropyl-B-D-thiogalacto pyranoside) is then added to a final concentration of 1 mM. IPTG induces by inactivating the lacI repressor, clearing the P/O leading to increased gene expression. Cells are grown an extra 3 to 4 hours. Cells are then harvested by centrifugation (20 mins at 6000Xg). The cell pellet is solubilized in the chaotropic agent 6 Molar Guanidine HCl. After clarification, solubilized stanniocalcin is purified from this solution by chromatography on a Nickel-Chelate column under conditions that allow for tight binding by proteins containing the 6-His tag (Hochuli, E. et al., Genetic Engineering, Principles & Methods, 12:87-98 (1990)). Protein renaturation out of GnHCl can be accomplished by several protocols (Jaenicke, R. and Rudolph, R., Protein Structure - A Practical Approach, IRL Press, New York (1990)). Initially, step dialysis is utilized to remove the GnHCL. Alternatively, the purified protein isolated from the Ni-chelate column can be bound to a second column

over which a decreasing linear GnHCL gradient is run. The protein is allowed to renature while bound to the column and is subsequently eluted with a buffer containing 250 mM Imidazole, 150 mM NaCl, 25 mM Tris-HCl pH 7.5 and 10% Glycerol. Finally, soluble protein is dialyzed against a storage buffer containing 5 mM Ammonium Bicarbonate. The purified protein was analyzed by SDS-PAGE.

#### Example 7

#### Method for determination of mutation of hMLH2 and hMLH3 in hereditary cancer

##### Isolation of Genomic Clones

A human genomic Pl library (Genomic Systems, Inc.) was screened by PCR using primers selected for the cDNA sequence of hMLH2 and hMLH3. Two clones were isolated for hMLH2 using primers 5' AAGCTGCTCTGTAAAAGCG 3' (SEQ ID No. 52) and 5' GCACCAGCATCCAAGGAG 3' (SEQ ID No. 53) and resulting in a 133 bp product. Three clones were isolated for hMLH3, using primers 5' CAACCATGAGACACATCGC 3' (SEQ ID No. 54) and 5' AGGTTAGTGAAGACTCTGTC 3' (SEQ ID No. 55) resulting in a 121 bp product. Genomic clones were nick-translated with digoxigenin deoxy-uridine 5'-triphosphate (Boehringer Mannheim), and FISH was performed as described (Johnson, Cg. et al., Methods Cell Biol., 35:73-99 (1991)). Hybridization with the hMLH3 probe were carried out using a vast excess of human cot-1 DNA for specific hybridization to the expressed hMLH3 locus. Chromosomes were counterstained with 4,6-diamino-2-phenylidole and propidium



iodide, producing a combination of C- and R-bands. Aligned images for precise mapping were obtained using a triple-band filter set (Chroma Technology, Brattleboro, VT) in combination with a cooled charge-coupled device camera (Photometrics, Tucson, AZ) and variable excitation wavelength filters (Johnson, Cv. et al., Genet. Anal. Tech. Appl., 8:75 (1991)). Image collection, analysis and chromosomal fractional length measurements were done using the ISee Graphical Program System (Inovision Corporation, Durham, NC).

#### Transcription coupled Translation Mutation Analysis

For purposes of IVSP analysis the hMLH2 gene was divided into three overlapping segments. The first segment included codons 1 to 500, while the middle segment included codons 270 to 755, and the last segment included codons 485 to the translational termination site at codon 933. The primers for the first segment were 5' GGATCCTAATACGACTCACT ATAGGGAGACCACCATGGAACAATTGCCTGCGG 3' (SEQ ID No. 56) and 5' CCTGCTCCACTCATCTGC 3' (SEQ ID No. 57), for the middle segment were 5' GGATCCTAATACGACTCACTATAGGGAGACCACCATGGAAGATATCTTAAAGTTAATCCG 3' (SEQ ID No. 58) and 5' GGCTTCTTCTACTCTATATGG 3' (SEQ ID No. 59), and for the final segment were 5' GGATCCTAATACGACTCACTATAGGGAGACCACCATGGCAGGTCTTGAAAACCTCTTCG 3' (SEQ ID No. 60) and 5' AAAACAAGTCAGTGAATCCTC 3' (SEQ ID No. 61). The primers used for mapping the stop mutation in patient CW all used the same 5' primer as the

first segment. The 3' nested primers were: 5' AAGCACATCTGTTTCTGCTG 3' (SEQ ID No. 62) codons 1 to 369; 5' ACGAGTAGATTCCCTTTAGGC 3' (SEQ ID No. 63) codons 1 to 290; and 5' CAGAACTGACATGAGAGCC 3' (SEQ ID No. 64) codons 1 to 214.

For analysis of hMLH3, the hMLH3 cDNA was amplified as a full-length product or as two overlapping segments. The primers for full-length hMLH3 were 5' GGATCCTAATACGACTCACTATAGGGAGACCACCATGGAGCGAGCTGAGAGC 3' (SEQ ID No. 65) and 5' AGGTTAGTGAAGACTCTGTC 3' (SEQ ID No. 66) (codons 1 to 863). For segment 1, the sense primer was the same as above and the antisense primer was 5' CTGAGGTCTCAGCAGGC 3' (SEQ ID No. 67) (codons 1 to 472). Segment 2 primers were 5' GGATCCTAATACGACTCACTATAGGGAGACCACCATGGTGTCATTTCAGACTGCG 3' (SEQ ID No. 68) and 5' AGGTTAGTGAAGACTCTGTC 3' (SEQ ID No. 69) (codons 415 to 863). Amplifications were done as described below.

The PCR products contained recognition signals for transcription by T7 RNA polymerase and for the initiation of translation at their 5' ends. PCR products were used as templates in coupled transcription-translation reactions containing 40 uCi of <sup>35</sup>S-methionine (NEN, Dupont). Samples were diluted in SDS sample buffer, and analyzed by electrophoresis on SDS-polyacrylamide gels containing a gradient of 10 to 20% acrylamide. The gels were fixed, treated with EnHance (Dupont), dried and subjected to autoradiography.

### RT-PCR and Direct Sequencing of PCR Products

cDNAs were generated from RNA of lymphoblastoid or tumor cells with Superscript II (Life Technologies). The cDNAs were then used as templates for PCR. The conditions for all amplifications were 35 cycles at 95°C for 30s, 52°C to 62°C for 60 to 120s, and 70°C for 60 to 120s, in buffer. The PCR products were directly sequenced and cloned into the T-tailed cloning vector PCR2000 (Invitrogen) and sequenced with T7 polymerase (United States Biochemical). For the direct sequencing of PCR products, PCR reactions were first phenolchloroform extracted and ethanol precipitated. Templates were directly sequenced using Sequitherm polymerase (Epicentre Technologies) and gamma-<sup>32</sup>P labelled primers as described by the manufacturer.

### Intron/Exon Boundaries and Genomic Analysis of Mutations

Intron/exon borders were determined by cycle-sequencing P1 clones using gamma-<sup>32</sup>P end labelled primers and SequiTherm polymerase as described by the manufacturer. The primers used to amplify the hMLH2 exon containing codons 195 to 233 were 5' TTATTTGGCAGAAAAGCAGAG (SEQ ID No. 70) 3' and 5' TTAAAAGACTAACCTCTTGCC 3' (SEQ ID No. 71), which produced a 215 bp product. The product was cycle sequenced using the primer 5' CTGCTGTTATGAACAATATGG 3' (SEQ ID No. 72). The primers used to analyze the genomic deletion of hMLH3 in patient GC were: for the 5' region

amplification 5' CAGAAGCAGTTGCAAAGCC 3' (SEQ ID No. 73) and 5' AAACCGTACTCTTCACACAC 3' (SEQ ID No. 74) which produces a 74 bp product containing codons 233 to 257, primers 5' GAGGAAAAGCTTTTGTGTC 3' (SEQ ID No. 75) and 5' CAGTGGCTGCTGACTGAC 3' (SEQ ID No. 76) which produce a 93 bp product containing the codons 347 to 377, and primers 5' TCCAGAACCAAGAAGGAGC 3' (SEQ ID No. 77) and 5' TGAGGTCTCAGCAGGC 3' (SEQ ID No. 78) which produce a 99 bp product containing the codons 439 to 472 of hMLH3.

TABLE 5

Summary of Mutations in HMLH2 and HMLH3  
from patients affected with HNPCC

Sample	Codon	Nucleotides	cDNA Change	Genomic Change	Predicted Coding Change
<u>HMLH2</u>					
CW	233		Skipped Exon	CAG to TAG	GLN to Stop Codon
<u>HMLH3</u>					
MM, NS, TF	20		CGG to CAG	CGG to CAG	ARG to GLN
GC	268 to 669		1,203 bp Deletion	Deletion	In-frame deletion
GCx	268 to 669		1,203 bp Deletion	Deletion	Frameshift, truncation

Numerous modifications and variations of the present invention are possible in light of the above teachings and, therefore, within the scope of the appended claims, the invention may be practiced otherwise than as particularly described.

## SEQUENCE LISTING

- (1) GENERAL INFORMATION:
- (i) APPLICANT: HUMAN GENOME SCIENCES, INC.
  - (ii) TITLE OF INVENTION: Human DNA Mismatch Repair Proteins
  - (iii) NUMBER OF SEQUENCES: 78
  - (iv) CORRESPONDENCE ADDRESS:
    - (A) ADDRESSEE: CARELLA, BYRNE, BAIN, GILFILLAN, CECCHI, STEWART & OLSTEIN
    - (B) STREET: 6 BECKER FARM ROAD
    - (C) CITY: ROSELAND
    - (D) STATE: NEW JERSEY
    - (E) COUNTRY: USA
    - (F) ZIP: 07068
  - (v) COMPUTER READABLE FORM:
    - (A) MEDIUM TYPE: 3.5 INCH DISKETTE
    - (B) COMPUTER: IBM PS/2
    - (C) OPERATING SYSTEM: MS-DOS
    - (D) SOFTWARE: WORD PERFECT 5.1
  - (vi) CURRENT APPLICATION DATA:
    - (A) APPLICATION NUMBER: PCT/US95/01035
    - (B) FILING DATE: 25 JAN 1995
    - (C) CLASSIFICATION: UNASSIGNED
  - (v) PRIOR APPLICATION DATA:
    - (A) APPLICATION NUMBER: 08/294,312
    - (B) FILING DATE: 23 AUG 1994
    - (C) CLASSIFICATION:
  - (vi) PRIOR APPLICATION DATA:
    - (A) APPLICATION NUMBER: 08/210,143
    - (B) FILING DATE: 16 MARCH 1994
    - (C) CLASSIFICATION:
  - (vii) PRIOR APPLICATION DATA:
    - (A) APPLICATION NUMBER: 08/187,757
    - (B) FILING DATE: 27 JAN 1994
    - (C) CLASSIFICATION:
  - (vi) ATTORNEY/AGENT INFORMATION:
    - (A) NAME: FERRARO, GREGORY D.
    - (B) REGISTRATION NUMBER: 36,134
    - (C) REFERENCE/DOCKET NUMBER: 325800-303
  - (viii) TELECOMMUNICATION INFORMATION:

(A) TELEPHONE: 201-994-1700  
(B) TELEFAX: 201-994-1744

## (2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 2525 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

GTGGAACATC	TAGACGTTTC	CTTGGCTCTT	CTGGCGCCAA	AATGTCGTTC	GTGGCAGGGG	60
TTATTCGGCG	GCTGGACGAG	ACAGTGGTGA	ACCGCATCGC	GGCGGGGGAA	GTTATCCAGC	120
GGCCAGCTAA	TGCTATCAAA	GAGATGATTG	AGAACTGTTT	AGATGCAAAA	TCCACAAGTA	180
TTCAAGTGAT	TGTTAAAGAG	GGAGGCCTGA	AGTTGATTCA	GATCCAAGAC	AATGGCACCG	240
GGATCAGGAA	AGAAGATCTG	GATATTGTAT	GTGAAAGTGT	CACTACTAGT	AAACTGCAGT	300
CCTTTGAGGA	TTTAGCCAGT	ATTTCTATCT	ATGGCTTTTC	AGGTGAGGCT	TTGGCCAGCA	360
TAAGCCATGT	GGCTCATGTT	ACTATTACAA	CGAAACAGC	TGATGGAAAG	TGTGCATACA	420
GAGCAAGTTA	CTCAGATGGA	AAACTGAAAG	CCCCTCCTAA	ACCATGTGCT	GGCAATCAAG	480
GGACCCAGAT	CACGGTGGAG	GACCTTTTTC	ACAACATAGC	CACGAGGAGA	AAAGCTTTAA	540
AAAATCCAAG	TGAAGAATAT	GGGAAAATTT	TGGAAGTTGT	TGGCAGGTAT	TCAGTACACA	600
ATGCAGGCAT	TAGTTTCTCA	GTAAAAAAC	AAGGAGAGAC	AGTAGCTGAT	GTTAGGACAC	660
TACCCAATGC	CTCAACCGTG	GACAATATTC	GCTCCGTCTT	GGGAAATGCT	GTTAGTCGAG	720
AACTGATAGA	AATTGGATGT	GAGGATAAAA	CCCTAGCCTT	CAAAATGAAT	GGTTACATAT	780
CCAATGCAAA	CTACTCAGTG	AAGAAGTGCA	TCTTCTTACT	CTTCATCAAC	CATCGTCTGG	840
TAGAATCAAC	TTCCCTTGAGA	AAAGCCATAG	AAACAGTGTA	TGCAGCCTAT	TTGCCAAAAA	900
ACACACACCC	ATTCCTGTAC	CTCAGTTTAG	AAATCAGTCC	CCAGAATGTG	GATGTTAATG	960
TGAACCCAC	AAAGCATGAA	GTTCACTTCC	TGCACGAGGA	GAGCATCCTG	GAGCGGGTGC	1020
AGCAGCACAT	CGAGAGCAAG	CTCCTGGGCT	CCAAATCCTC	CAGGATGTAC	TTCAACCAGA	1080
CTTTGCTACC	AGGACTTGCT	GGCCCTCTG	GGGAGATGGT	TAAATCCACA	ACAAGTCTCA	1140
CCTCGTCTTC	TACTTCTGGA	AGTAGTGATA	AGGTCTATGC	CCACCAGATG	GTTCTGTACAG	1200
ATTCCTGGGA	ACAGAAGCTT	GATGCATTTC	TGCAGCCTCT	GAGCAAACCC	CTGTCCAGTC	1260
AGCCCCAGGC	CATTGTCAAC	GAGGATAAGA	CAGATATTTT	TAGTGGCAGG	GCTAGGCAGC	1320
AAGATGAGGA	GATGCTTGAA	CTCCAGCCCC	CTGCTGAAGT	GGCTGCCAAA	AATCAGAGCT	1380
TGGAGGGGGA	TACAACAAAG	GGGACTTCAG	AAATGTCAGA	GAAGAGAGGA	CCTACTTCCA	1440
GCAACCCAG	AAAGAGACAT	CGGGAAGATT	CTGATCTCCA	AATCCTCGAA	GATGATTCCC	1500
GAAAGGAAAT	GACTGCAGCT	TGTACCCCCC	GGAGAAGGAT	CATTAACTTC	ACTAGTGTTT	1560
TGAGTCTCCA	GGAAGAAATT	AATGAGCAGG	GACATGAGGT	TCTCCGGGAG	ATGTTGCATA	1620
ACCACCTCTT	CGTGGGCTGT	GTGAATCCTC	AGTGGGCCTT	GGCACAGCAT	CAAACCAAGT	1680
TATACCTTCT	CAACACCACC	AAGCTTAGTG	AAGAACTGTT	CTACCAGATA	CTCATTTATG	1740
ATTTTGCCAA	TTTTGGTGTT	CTCAGGTTAT	CGGAGCCAGC	ACCGCTCTTT	GACCTTGCCA	1800
TGCTTCCCTT	ACATAGTCCA	GAGAGTGGCT	GGACAGAGGA	AGATGGTCCC	AAAGAAGGAC	1860
CTTTGGAAAT	TGATGAGGAA	GGGAACCTGA	TTGGATTACC	CCTTCTGATT	GACAACTATG	1920
TGCCCCCTTT	GGAGGGACTG	CCTATCTTCA	TTCTTCCACT	AGCCACTGAG	GTGAATTGGG	1980
ACGAAGAAAA	GGAATGTTTT	GAAAGCCTCA	GTAAGAATG	CGCTATGTTT	TATTCCATCC	2040
GGAAGCAGTA	CATATCTGAG	GAGTCGACCC	TCTCAGGCCA	GCAGAGTGAA	GTGCCTGGCT	2100
CCATTCCAAA	CTCCTGGAAG	TGGACTGTGG	AACACATTGT	CTATAAAGCC	TTGCGCTCAC	2160
ACATTCTGCC	TCCTAAACAT	TCCACAGAAG	ATGGAAATAT	CCTGCAGCTT	GCTAACCTGC	2220
CTGATCTATA	CAAAGTCTTT	GAGAGGTGTT	AAATATGGTT	ATTTATGCAC	TGTGGGATGT	2280
GTTCTTCTTT	CTCTGTATTC	CGATACAAAG	TGTTGTACTA	AAGTGTGATA	TACAAAGTGT	2340
ACCAACATAA	GTGTTGGTAG	CACCTTAAGAC	TTATACTTGC	CTTCTGATAG	TATTCTTTTA	2400
TACACAGTGG	ATTGATTATA	AATAAATAGA	TGTGCTTTAA	CATAAAAAAA	AAAAAAAAAA	2460
AAAAA						2520
						2525

## (2) INFORMATION FOR SEQ ID NO:2:

- (i) SEQUENCE CHARACTERISTICS  
 (A) LENGTH: 756 AMINO ACIDS  
 (B) TYPE: AMINO ACID  
 (C) STRANDEDNESS:  
 (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: PROTEIN

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

Met	Ser	Phe	Val	Ala	Gly	Val	Ile	Arg	Arg	Leu	Asp	Glu	Thr	Val	5	10	15
Val	Asn	Arg	Ile	Ala	Gly	Glu	Val	Ile	Gln	Arg	Pro	Ala	Asn	20	25	30	
Ala	Ile	Lys	Glu	Met	Ile	Glu	Asn	Cys	Leu	Asp	Ala	Lys	Ser	Thr	35	40	45
Ser	Ile	Gln	Val	Ile	Val	Lys	Glu	Gly	Gly	Leu	Lys	Leu	Ile	Gln	50	55	60
Ile	Gln	Asp	Asn	Gly	Thr	Gly	Ile	Arg	Lys	Glu	Asp	Leu	Asp	Ile	65	70	75
Val	Cys	Glu	Arg	Phe	Thr	Thr	Ser	Lys	Leu	Gln	Ser	Phe	Glu	Asp	80	85	90
Leu	Ala	Ser	Ile	Ser	Thr	Tyr	Gly	Phe	Arg	Gly	Glu	Ala	Leu	Ala	95	100	105
Ser	Ile	Ser	His	Val	Ala	His	Val	Thr	Ile	Thr	Thr	Lys	Thr	Ala	110	115	120
Asp	Gly	Lys	Cys	Ala	Tyr	Arg	Ala	Ser	Tyr	Ser	Asp	Gly	Lys	Leu	125	130	135
Lys	Ala	Pro	Pro	Lys	Pro	Cys	Ala	Gly	Asn	Gln	Gly	Thr	Gln	Ile	140	145	150
Thr	Val	Glu	Asp	Leu	Phe	Tyr	Asn	Ile	Ala	Thr	Arg	Arg	Lys	Ala	155	160	165
Leu	Lys	Asn	Pro	Ser	Glu	Glu	Tyr	Gly	Lys	Ile	Leu	Glu	Val	Val	170	175	180
Gly	Arg	Tyr	Ser	Val	His	Asn	Ala	Gly	Ile	Ser	Phe	Ser	Val	Lys	185	190	195
Lys	Gln	Gly	Glu	Thr	Val	Ala	Asp	Val	Arg	Thr	Leu	Pro	Asn	Ala	200	205	210
Ser	Thr	Val	Asp	Asn	Ile	Arg	Ser	Val	Phe	Gly	Asn	Ala	Val	Ser	215	220	225
Arg	Glu	Leu	Ile	Glu	Ile	Gly	Cys	Glu	Asp	Lys	Thr	Leu	Ala	Phe	230	235	240
Lys	Met	Asn	Gly	Tyr	Ile	Ser	Asn	Ala	Asn	Tyr	Ser	Val	Lys	Lys	245	250	255
Cys	Ile	Phe	Leu	Leu	Phe	Ile	Asn	His	Arg	Leu	Val	Glu	Ser	Thr	260	265	270
Ser	Leu	Arg	Lys	Ala	Ile	Glu	Thr	Val	Tyr	Ala	Ala	Tyr	Leu	Pro	275	280	285
Lys	Asn	Thr	His	Pro	Phe	Leu	Tyr	Leu	Ser	Leu	Glu	Ile	Ser	Pro	290	295	300
Gln	Asn	Val	Asp	Val	Asn	Val	His	Pro	Thr	Lys	His	Glu	Val	His			



Phe	Leu	His	Glu	305	Glu	Ser	Ile	Leu	Glu	310	Arg	Val	Gln	Gln	His	Ile	315
				320						325							330
Glu	Ser	Lys	Leu	335	Leu	Gly	Ser	Asn	Ser	340	Ser	Arg	Met	Tyr	Phe	Thr	345
Gln	Thr	Leu	Leu	350	Pro	Gly	Leu	Ala	Ala	355	Pro	Ser	Gly	Glu	Met	Val	360
Lys	Ser	Thr	Thr	365	Ser	Leu	Thr	Ser	Ser	370	Ser	Thr	Ser	Gly	Ser	Ser	375
Asp	Lys	Val	Tyr	380	Ala	His	Gln	Met	Val	385	Arg	Thr	Asp	Ser	Arg	Glu	390
Gln	Lys	Leu	Asp	395	Ala	Phe	Leu	Gln	Pro	400	Leu	Ser	Lys	Pro	Leu	Ser	405
Ser	Gln	Pro	Gln	410	Ala	Ile	Val	Thr	Glu	415	Asp	Lys	Thr	Asp	Ile	Ser	420
Ser	Gly	Arg	Ala	425	Arg	Gln	Gln	Asp	Glu	430	Glu	Met	Leu	Glu	Leu	Pro	435
Ala	Pro	Ala	Glu	440	Val	Ala	Ala	Lys	Asn	445	Gln	Ser	Leu	Glu	Gly	Asp	450
Thr	Thr	Lys	Gly	455	Thr	Ser	Glu	Met	Ser	460	Glu	Lys	Arg	Gly	Pro	Thr	465
Ser	Ser	Asn	Pro	470	Arg	Lys	Arg	His	Arg	475	Glu	Asp	Ser	Asp	Val	Glu	480
Met	Val	Glu	Asp	485	Asp	Ser	Arg	Lys	Glu	490	Met	Thr	Ala	Ala	Cys	Thr	495
Pro	Arg	Arg	Arg	500	Ile	Ile	Asn	Leu	Thr	505	Ser	Val	Leu	Ser	Leu	Gln	510
Glu	Glu	Ile	Asn	515	Glu	Gln	Gly	His	Glu	520	Val	Leu	Arg	Glu	Met	Leu	525
His	Asn	His	Ser	530	Phe	Val	Gly	Cys	Val	535	Asn	Pro	Gln	Trp	Ala	Leu	540
Ala	Gln	His	Gln	545	Thr	Lys	Leu	Tyr	Leu	550	Leu	Asn	Thr	Thr	Lys	Leu	555
Ser	Glu	Glu	Leu	560	Phe	Tyr	Gln	Ile	Leu	565	Ile	Tyr	Asp	Phe	Ala	Asn	570
Phe	Gly	Val	Leu	575	Arg	Leu	Ser	Glu	Pro	580	Ala	Pro	Leu	Phe	Asp	Leu	585
Ala	Met	Leu	Ala	590	Leu	Asp	Ser	Pro	Glu	595	Ser	Gly	Trp	Thr	Glu	Glu	600
Asp	Gly	Pro	Lys	605	Glu	Gly	Leu	Ala	Glu	610	Tyr	Ile	Val	Glu	Phe	Leu	615
Lys	Lys	Lys	Ala	620	Glu	Met	Leu	Ala	Asp	625	Tyr	Phe	Ser	Leu	Glu	Ile	630
Asp	Glu	Glu	Gly	635	Asn	Leu	Ile	Gly	Leu	640	Pro	Leu	Leu	Thr	Asp	Asn	645
Tyr	Val	Pro	Pro	650	Leu	Glu	Gly	Leu	Pro	655	Ile	Phe	Ile	Leu	Arg	Leu	660
Ala	Thr	Glu	Val	665	Asn	Trp	Asp	Glu	Glu	670	Lys	Glu	Cys	Phe	Glu	Ser	675
Leu	Ser	Lys	Glu	680	Cys	Ala	Met	Phe	Tyr	685	Ser	Ile	Arg	Lys	Gln	Tyr	690
Ile	Ser	Glu	Glu		Ser	Thr	Leu	Ser	Gly		Gln	Gln	Ser	Glu	Val	Pro	

	695		700		705
Gly Ser Ile Pro Asn Ser Trp Lys Trp		Thr Val Glu His Ile Val			
	710		715		720
Tyr Lys Ala Leu Arg Ser His Ile Leu		Pro Pro Lys His Phe Thr			
	725		730		735
Glu Asp Gly Asn Ile Leu Gln Leu Ala		Asn Leu Pro Asp Leu Tyr			
	740		745		750
Lys Val Phe Glu Arg Cys					
	755				

## (2) INFORMATION FOR SEQ ID NO:3:

## (i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 3063 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

## (ii) MOLECULE TYPE: cDNA

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

GGCAGCAGTG	GCTGCTTGCG	GCTAGTGGAT	GGTAATTGCC	TGCCTCGCGC	TAGCAGCAAG	60
CTGCTCTGTT	AAAAGCGAAA	ATGAAACAAT	TGCCTGCGGC	AACAGTTCGA	CTCCTTTCAA	120
GTTCTCAGAT	CATCACTTCG	GTGGTCAGTG	TTGTAAAGA	GCTTATTGAA	AATCCTTGG	180
ATGCTGGTGC	CACAAGCGTA	GATGTAAAC	TGGAGAACTA	TGGATTGTAT	AAAATTGAGG	240
TGCGAGATAA	CGGGGAGGGT	ATCAAGGCTG	TTGATGCACC	TGTAATGGCA	ATGAAGTACT	300
ACACCTCAAA	AATAAATAGT	CATGAAGATC	TTGAAAATTT	GACAACTTAC	GGTTTTCGTG	360
GAGAAGCCTT	GGGGTCAATT	TGTTGTATAG	CTGAGGTTTT	AATTACAACA	AGAACGGCTG	420
CTGATAATTT	TAGCACCCAG	TATGTTTTAG	ATGGCAGTGG	CCACATACTT	TCTCAGAAAC	480
CTTCACATCT	TGGTCAAGGT	ACAACGTGTA	CTGCTTTAAG	ATTATTAAAG	AATCTACCTG	540
TAAGAAAGCA	GTTTTACTCA	ACTGCAAAAA	AATGTAAAGA	TGAAATAAAA	AAGATCCAAG	600
ATCTCCTCAT	GAGCTTTGGT	ATCCTTAAAC	CTGACTTAAG	GATTGTCTTT	GTACATAACA	660
AGGCAGTTAT	TTGGCAGAAA	AGCAGAGTAT	CAGATCACAA	GATGGCTCTC	ATGTCAGTTC	720
TGGGACTCTG	TGTTATGAAC	AATATGGAAT	CCTTTCAGTA	CCACTCTGAA	GAATCTCAGA	780
TTTATCTCAG	TGGATTTCTT	CCAAAGTGTG	ATGCAGACCA	CTCTTTCACT	AGTCTTTCAA	840
CACCAGAAAG	AAGTTTCATC	TTCAATAACA	GTCCAGCAGT	ACATCAAAAA	GATATCTTAA	900
AGTTAATCCG	ACATCATTAC	AATCTGAAAT	GCCTAAAGGA	ATCTACTCGT	TTGTATCCTG	960
TTTTCTTTCT	GAAAATCGAT	GTTCCCTACAG	CTGATGTTGA	TGTAAATTTA	ACACCAGATA	1020
AAAGCCAAGT	ATTATTACAA	AATAAGGAAT	CTGTTTTAAT	TGCTCTTGAA	AATCTGATGA	1080
CGACTTGTTA	TGGACCATTA	CCTAGTACAA	ATTCTTATGA	AAATAATAAA	ACAGATGTTT	1140
CCGCAGCTGA	CATCGTTCCT	AGTAAAACAG	CAGAAAACAGA	TGTGCTTTTT	AATAAAGTGG	1200
AATCATCTGG	AAAGAATTAT	TCAAATGTTG	ATACTTCAGT	CATTCCATTG	CAAAATGATA	1260
TGCATAATGA	TGAATCTGGA	AAAAACACTG	ATGATTGTTT	AAATCACCAG	ATAAGTATTG	1320
GTGACTTTGG	TTAGTTCAT	TGTAGTAGTG	AAATTTCTAA	CATTGATAAA	AACACTAAGA	1380
ATGCATTTCA	GGACATTTCA	ATGAGTAATG	TATCATGGGA	GAACCTCTAG	ACGGAATATA	1440
GTAAACTTGT	TTTTATAAGT	TCCGTTAAGC	ACACCCAGTC	AGAAAATGGC	AATAAAGACC	1500
ATATAGATGA	GAGTGGGGAA	AATGAGGAAG	AAGCAGGTCT	TGAAAATCT	TCGGAAATTT	1560
CTGCAGATGA	GTGGAGCAGG	GGAAATATAC	TTAAAAATTC	AGTGGGAGAG	AATATTGAAC	1620
CTGTGAAAAT	TTTAGTGCCT	GAAAAAAGTT	TACCATGTAA	AGTAAGTAAT	AATAATTATC	1680
CAATCCCTGA	ACAAATGAAT	CTTAATGAAG	ATTCATGTAA	CAAAAAATCA	AATGTAATAG	1740
ATAATAAATC	TGGAAAAGTT	ACAGCTTATG	ATTACTTAG	CAATCGAGTA	ATCAAGAAAC	1800
CCATGTCAGC	AAGTGCTCTT	TTTGTTCAGG	ATCATCGTCC	TCAGTTTCTC	ATAGAAAATC	1860
CTAAGACTAG	TTTAGAGGAT	GCAACACTAC	AAATTGAAGA	ACTGTGGAAG	ACATTGAGTG	1920
AAGAGGAAAA	ACTGAAATAT	GAAGAGAAGG	CTACTAAAGA	CTTGGNACGA	TACAATAGTC	1980
AAATGAAGAG	AGCCATTGAA	CAGGAGTCAC	AAATGTCACT	AAAAGATGGC	AGAAAAAAGA	2040
TAAACCCAC	CAGCGCATGG	AATTTGGCCC	AGAAGCACAA	GTTAAAAACC	TCATTATCTA	2100
ATCAACCANA	ACTTGATGAA	CTCCTTCAGT	CCCAAATTGA	AAAAAGAAGG	AGTCAAAATA	2160
TTAAAATGGT	ACAGATCCCC	TTTTCTATGA	AAACTTAA	AATAAATTTT	AAGAAACAAA	2220

ACAAAGTTGA	CTTAGAAGAG	AAGGATGAAC	CTTGCTTGAT	CCACAATCTC	AGGTTTCCTG	2280
ATGCATGGCT	AATGACATCC	AAAACAGAGG	TAATGTTATT	AAATCCATAT	AGAGTAGAAG	2340
AAGCCCTGCT	ATTTAAAAGA	CTTCTTGAGA	ATCATAAACT	TCCTGCAGAG	CCACTGGAAA	2400
AGCCAATTAT	GTTAACAGAG	AGTCTTTTTA	ATGGATCTCA	TTATTTAGAC	GTTTTATATA	2460
AAATGACAGC	AGATGACCAA	AGATACAGTG	GATCAACTTA	CCTGCTGAT	CCTCGTCTTA	2520
CAGCGAATGG	TTTCAAGATA	AAATTGATAC	CAGGAGTTTC	AATTACTGAA	AATTACTTGG	2580
AAATAGAAGG	AATGGCTAAT	TGTCTCCCAT	TCTATGGAGT	AGCAGATTTA	AAAGAAATTC	2640
TTAATGCTAT	ATTAAACAGA	AATGCAAAGG	AAGTTTATGA	ATGTAGACCT	CGCAAAGTGA	2700
TAAGTTATTT	AGAGGGAGAA	GCAGTGCCTC	TATCCAGACA	ATTACCCATG	TACTTATCAA	2760
AAGAGGACAT	CCAAGACATT	ATCTACAGAA	TGAAGCACCA	GTTTGGAAAT	GAAATTAAGG	2820
AGTGTTGTTCA	TGGTCGCCCA	TTTTTTCATC	ATTTAACCTA	TCTTCCAGAA	ACTACATGAT	2880
TAAATATGTT	TAAGAAGATT	AGTTACCATT	GAAATGGTT	CTGTCATAAA	ACAGCATGAG	2940
TCTGGTTTTA	AATTATCTTT	GTATTATGTG	TCACATGGTT	ATTTTTTAAA	TGAGGATTCA	3000
CTGACTTGTT	TTTATATTGA	AAAAAGTTCC	ACGTATTGTA	GAAAACGTAA	ATAACTAAT	3060
AAC						3063

## (2) INFORMATION FOR SEQ ID NO:4:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 931 BASE PAIRS  
 (B) TYPE: AMINO ACID  
 (C) STRANDEDNESS:  
 (D) TOPOLOGY: LINEAR

## (ii) MOLECULE TYPE: PROTEIN (XI)

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

Met	Lys	Gln	Leu	Pro	Ala	Ala	Thr	Val	Arg	Leu	Leu	Ser	Ser	Ser
			5						10					15
Gln	Ile	Ile	Thr	Ser	Val	Val	Ser	Val	Val	Lys	Glu	Leu	Ile	Glu
			20						25					30
Asn	Ser	Leu	Asp	Ala	Gly	Ala	Thr	Ser	Val	Asp	Val	Lys	Leu	Glu
			35						40					45
Asn	Tyr	Gly	Phe	Asp	Lys	Ile	Glu	Val	Arg	Asp	Asn	Gly	Glu	Gly
			50						55					60
Ile	Lys	Ala	Val	Asp	Ala	Pro	Val	Met	Ala	Met	Lys	Tyr	Tyr	Thr
			65						70					75
Ser	Lys	Ile	Asn	Ser	His	Gly	Asp	Leu	Glu	Asn	Leu	Thr	Thr	Tyr
			80						85					90
Gly	Phe	Arg	Gly	Glu	Ala	Leu	Gly	Ser	Ile	Cys	Cys	Ile	Ala	Glu
			95						100					105
Val	Leu	Ile	Thr	Thr	Arg	Thr	Ala	Ala	Asp	Asn	Phe	Ser	Thr	Gln
			110						115					120
Tyr	Val	Leu	Asp	Gly	Ser	Gly	His	Ile	Leu	Ser	Gln	Lys	Pro	Ser
			125						130					135
His	Leu	Gly	Gln	Gly	Thr	Thr	Val	Thr	Ala	Leu	Arg	Leu	Phe	Lys
			140						145					150
Asn	Leu	Pro	Val	Arg	Lys	Gln	Phe	Tyr	Ser	Thr	Ala	Lys	Lys	Cys
			155						160					165
Lys	Asp	Glu	Ile	Lys	Lys	Ile	Gln	Asp	Leu	Leu	Met	Ser	Phe	Gly
			170						175					180
Ile	Leu	Lys	Pro	Asp	Leu	Arg	Ile	Val	Phe	Val	His	Asn	Lys	Ala
			185						190					195
Val	Ile	Trp	Gln	Lys	Ser	Arg	Val	Ser	Asp	His	Lys	Met	Ala	Leu

Met Ser Val Leu	200	Met Asn Asn Met Glu Ser	205	210
Gly Thr Ala Val	215	Asn Asn Met Glu Ser	220	225
Gln Tyr His Ser	230	Tyr Leu Ser Gly Phe	235	240
Pro Lys Cys Asp	245	Thr Ser Leu Ser Thr	250	255
Glu Arg Ser Phe	260	Arg Pro Val His Gln	265	270
Asp Ile Leu Lys	275	Tyr Asn Leu Lys Cys	280	285
Lys Glu Ser Thr	290	Phe Phe Leu Lys Ile	295	300
Val Pro Thr Ala	305	Leu Thr Pro Asp Lys	310	315
Gln Val Leu Leu	320	Val Leu Ile Ala Leu	325	330
Asn Leu Met Thr	335	Leu Pro Ser Thr Asn	340	345
Tyr Glu Asn Asn	350	Ala Ala Asp Ile Val	355	360
Ser Lys Thr Ala	365	Phe Asn Lys Val Glu	370	375
Ser Gly Lys Asn	380	Thr Ser Val Ile Pro	385	390
Gln Asn Asp Met	395	Gly Lys Asn Thr Asp	400	405
Cys Leu Asn His	410	Asp Phe Gly Tyr Gly	415	420
Cys Ser Ser Glu	425	Lys Asn Thr Lys Asn	430	435
Phe Gln Asp Ile	440	Ser Trp Glu Asn Ser	445	450
Thr Glu Tyr Ser	455	Ser Ser Val Lys His	460	465
Gln Ser Glu Asn	470	Ile Asp Glu Ser Gly	475	480
Asn Glu Glu Glu	485	Ser Ser Glu Ile Ser	490	495
Asp Glu Trp Ser	500	Lys Asn Ser Val Gly	505	510
Asn Ile Glu Pro	515	Pro Glu Lys Ser Leu	520	525
Cys Lys Val Ser	530	Ile Pro Glu Gln Met	535	540
Leu Asn Glu Asp	545	Ser Asn Val Ile Asp	550	555
Lys Ser Gly Lys	560	Leu Leu Ser Asn Arg	565	570
Ile Lys Lys Pro	575	Leu Phe Val Gln Asp	580	585
Arg Pro Gln Phe		Lys Thr Ser Leu Glu		Asp

Ala Thr Leu Gln	590	Ile Glu Glu Leu Trp	595	Lys Thr Leu Ser Glu	600
Glu Lys Leu Lys	605	Tyr Glu Glu Lys Ala	610	Thr Lys Asp Leu Xaa	615
Tyr Asn Ser Gln	620	Met Lys Arg Ala Ile	625	Glu Gln Glu Ser Gln	630
Ser Leu Lys Asp	635	Gly Arg Lys Lys Ile	640	Lys Pro Thr Ser Ala	645
Asn Leu Ala Gln	650	Lys His Lys Leu Lys	655	Thr Ser Leu Ser Asn	660
Pro Xaa Leu Asp	665	Glu Leu Leu Gln Ser	670	Gln Ile Glu Lys Arg	675
Ser Gln Asn Ile	680	Lys Met Val Gln Ile	685	Pro Phe Ser Met Lys	690
Leu Lys Ile Asn	695	Phe Lys Lys Gln Asn	700	Lys Val Asp Leu Glu	705
Lys Asp Glu Pro	710	Cys Leu Ile His Asn	715	Leu Arg Phe Pro Asp	720
Trp Leu Met Thr	725	Ser Lys Thr Glu Val	730	Met Leu Leu Asn Pro	735
Arg Val Glu Glu	740	Ala Leu Leu Phe Lys	745	Arg Leu Leu Glu Asn	750
Lys Leu Pro Ala	755	Glu Pro Leu Glu Lys	760	Arg Leu Leu Glu Asn	765
Ser Leu Phe Asn	770	Gly Ser His Tyr Leu	775	Pro Ile Met Leu Thr	780
Thr Ala Asp Asp	785	Gln Arg Tyr Ser Gly	790	Asp Val Leu Tyr Lys	795
Pro Arg Leu Thr	800	Ala Asn Gly Phe Lys	805	Ser Thr Tyr Leu Ser	810
Val Ser Ile Thr	815	Glu Asn Tyr Leu Glu	820	Ile Lys Leu Ile Pro	825
Cys Leu Pro Phe	830	Tyr Gly Val Ala Asp	835	Ile Glu Gly Met Ala	840
Ala Ile Leu Asn	845	Arg Asn Ala Lys Glu	850	Leu Lys Glu Ile Leu	855
Arg Lys Val Ile	860	Ser Tyr Leu Glu Gly	865	Val Tyr Glu Cys Arg	870
Arg Gln Leu Pro	875	Met Tyr Leu Ser Lys	880	Glu Ala Val Arg Leu	885
Ile Tyr Arg Met	890	Lys His Gln Phe Gly	895	Glu Asp Ile Gln Asp	900
Val His Gly Arg	905	Pro Phe Phe His His	910	Asn Glu Ile Lys Glu	915
Thr	920	Leu Thr Tyr Leu Pro	925	Leu Thr Tyr Leu Pro	930

- (2) INFORMATION FOR SEQ ID NO:5:  
 (i) SEQUENCE CHARACTERISTICS  
 (A) LENGTH: 2771 BASE PAIRS  
 (B) TYPE: NUCLEIC ACID  
 (C) STRANDEDNESS: SINGLE

## (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

CGAGGCGGAT	CGGGTGTTC	ATCCATGGAG	CGAGCTGAGA	GCTCGAGTAC	AGAACCTGCT	60
AAGGCCATCA	AACCTATTGA	TCGGAAGTCA	GTCCATCAGA	TTTGCTCTGG	GCAGGTGGTA	120
CTGAGTCTAA	GCACTGCGGT	AAAGGAGTTA	GTAGAAAACA	GTCTGGATGC	TGGTGCCACT	180
AATATTGATC	TAAAGCTTAA	GGACTATGGA	GTGGATCTTA	TTGAAGTTTC	AGACAATGGA	240
TGTGGGGTAG	AAGAAGAAAA	CTTCGAAGGC	TTAACTCTGA	AACATCACAC	ATCTAAGATT	300
CAAGAGTTTG	CCGACCTAAC	TCAGGTTGAA	ACTTTTGGCT	TTCGGGGGGA	AGCTCTGAGC	360
TCACTTTGTG	CAGTGAGCGA	TGTCAACATT	TCTACCTGCC	ACGCATCGGC	GAAGGTTGGA	420
ACTCGACTGA	TGTTTGATCA	CAATGGGAAA	ATTATCCAGA	AAACCCCTTA	CCCCGCCCC	480
AGAGGGACCA	CAGTCAGCGT	GCAGCAGTTA	TTTTCCACAC	TACCTGTGCG	CCATAAGGAA	540
TTTCAAAGGA	ATATTAAGAA	GGAGTATGCC	AAAATGGTCC	AGGTCCTTACA	TGCATACTGT	600
ATCATTTCAG	CAGGCATCCG	TGTAAGTTGC	ACCAATCAGC	TTGGACAAGG	AAAACGACAG	660
CCTGTGGTAT	GCAAGGTGG	AAGCCCCAGC	ATAAAGGAAA	ATATCGGCTC	TGTGTTTGGG	720
CAGAAGCAGT	TGCAAAGCCT	CATTCCCTTT	GTTCAGCTGC	CCCCTAGTGA	CTCCGTGTGT	780
GAAGAGTACG	GTTTGAGCTG	TTCGGATGCT	CTGCATAATC	TTTTTTACAT	CTCAGGTTTC	840
ATTTCACAAT	GCACGCATGG	AGTTGGAAGG	AGTTCAACAG	ACAGACAGTT	TTTCTTTATC	900
AACCGGCGGG	CTTGTGACCC	AGCAAAGGTC	TGCAGACTCG	TGAATGAGGT	CTACCACATG	960
TATAATCGAC	ACCAGTATCC	ATTTGTTGTT	CTTAACATTT	CTGTTGATTTC	AGAATGCGTT	1020
GATATCAATG	TTACTCCAGA	TAAAAGGCAA	ATTTTGCTAC	AAGAGGAAAA	GCTTTTGTGT	1080
GCAGTTTTAA	AGACCTCTTT	GATAGGAATG	TTTGATAGTG	ATGTCACCAA	GCTAAATGTC	1140
AGTCAGCAGC	CACTGCTGGA	TGTTGAAGGT	AACTTAATAA	AAATGCATGC	AGCGGATTTG	1200
GAAAAGCCCA	TGGTAGAAAA	GCAGGATCAA	TCCCCTTCAT	TAAGGACTGG	AGAAAGAAAA	1260
AAAGACGTGT	CCATTTCCAG	ACTGCGAGAG	GCCTTTTCTC	TTCGTCACAC	AACAGAGAAC	1320
AAGCCTCACA	GCCCAAAGAC	TCCAGAACCA	AGAAGGAGCC	CTCTAGGACA	GAAAAGGGGT	1380
ATGCTGTCTT	CTAGCACTTC	AGGTGCCATC	TCTGACAAAG	GCGTCCTGAG	ACCTCAGAAA	1440
GAGGCAGTGA	GTTCCAGTCA	CGGACCCAGT	GACCCTACGG	ACAGAGCGGA	GGTGGAGAAG	1500
GACTCGGGGC	ACGGCAGCAC	TTCCGTGGAT	TCTGAGGGGT	TCAGCATCCC	AGACACGGGC	1560
AGTCACTGCA	GCAGCGAGTA	TGCGGCCAGC	TCCCAGGGG	ACAGGGGCTC	GCAGGAACAT	1620
GTGGACTCTC	AGGAGAAAGC	GCCTGAAACT	GACGACTCTT	TTTCAGATGT	GGACTGCCAT	1680
TCAAACCAGG	AAGATACCGG	ATGTAAATTT	CGAGTTTTCG	CTCAGCCAAC	TAATCTCGCA	1740
ACCCCAAACA	CAAAGCGTTT	TAAAAAGAA	GAAATCTTTT	CCAGTTCTGA	CATTTGTCAA	1800
AAGTTAGTAA	ATACTCAGGA	CATGTCAGCC	TCTCAGGTTG	ATGTAGCTGT	GAAAATTAAT	1860
AAGAAAGTTG	TGCCCCTGGA	CTTTTCTATG	AGTTCCTTAG	CTAAACGAAT	AAAGCAGTTA	1920
CATCATGAAG	CACAGCAAAG	TGAAGGGGAA	CAGAATTACA	GGAAGTTTAG	GGCAAAGATT	1980
TGTCCTGGAG	AAAATCAAGC	AGCCGAAGAT	GAACTAAGAA	AAGAGATAAG	TAAAACGATG	2040
TTTGAGAGAA	TGGAAATCAT	TGGTCAGTTT	AACCTGGGAT	TTATAATAAC	CACACTGAAT	2100
GAGGATATCT	TCATAGTGGA	CCAGCATGCC	ACGGACGAGA	AGTATAACTT	CGAGATGCTG	2160
CAGCAGCACA	CCGTGCTCCA	GGGGCAGACG	CTCATAGCAC	CTCAGACTCT	CAACTTAACT	2220
GCTGTTAATG	AAGCTGTTCT	GATAGAAAAT	CTGGAAATAT	TTAGAAAAGAA	TGGCTTTGAT	2280
TTTGTATATG	ATGAAAATGC	TCCAGTCACT	GAAAGGGCTA	AACTGATTTC	CTTGCCAACT	2340
AGTAAAAACT	GGACCTTCGG	ACCCACAGGAC	GTGATGAAC	TGATCTTCAT	GCTGAGCGAC	2400
AGCCCTGGGG	TCATGTGCCG	GCCTTCCCGA	GTCAAGCAGA	TGTTTGCCCTC	CAGAGCCTGC	2460
CGGAAGTCGG	TGATGATTGG	GACTGCTCTT	AACACAAGCG	AGATGAAGAA	ACTGATCACC	2520
CACATGGGGG	AGATGGACCA	CCCCTGGAAC	TGTCCCCATG	GAAGGCCAAC	CATGAGACAC	2580
ATCGCCAACC	TGGGTGTCAT	TTCTCAGAAC	TGACCGTAGT	CACGTATGGA	AATAATTGGT	2640
TTTATCGCAG	ATTTTATGT	TTTGAAAGAC	AGAGTCTTCA	CTAACCTTTT	TTGTTTTAAA	2700
ATGAAACCTG	CTACTTAAAA	AAAATACACA	TCACACCCAT	TAAAAGTGA	TCCTGAGAAC	2760
CTTTTCAAAC	C					2771

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 862 AMINO ACIDS

(B) TYPE: AMINO ACID

(C) STRANDEDNESS:

## (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: PROTEIN

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

Met	Glu	Arg	Ala	Glu	Ser	Ser	Ser	Thr	Glu	Pro	Ala	Lys	Ala	Ile	5	10	15
Lys	Pro	Ile	Asp	Arg	Lys	Ser	Val	His	Gln	Ile	Cys	Ser	Gly	Gln	20	25	30
Val	Val	Leu	Ser	Leu	Ser	Thr	Ala	Val	Lys	Glu	Leu	Val	Glu	Asn	35	40	45
Ser	Leu	Asp	Ala	Gly	Ala	Thr	Asn	Ile	Asp	Leu	Lys	Leu	Lys	Asp	50	55	60
Tyr	Gly	Val	Asp	Leu	Ile	Glu	Val	Ser	Asp	Asn	Gly	Cys	Gly	Val	65	70	75
Glu	Glu	Glu	Asn	Phe	Glu	Gly	Leu	Thr	Leu	Lys	His	His	Thr	Ser	80	85	90
Lys	Ile	Gln	Glu	Phe	Ala	Asp	Leu	Thr	Gln	Val	Glu	Thr	Phe	Gly	95	100	105
Phe	Arg	Gly	Glu	Ala	Leu	Ser	Ser	Leu	Cys	Ala	Leu	Ser	Asp	Val	110	115	120
Thr	Ile	Ser	Thr	Cys	His	Ala	Ser	Ala	Lys	Val	Gly	Thr	Arg	Leu	125	130	135
Met	Phe	Asp	His	Asn	Gly	Lys	Ile	Ile	Gln	Lys	Thr	Pro	Tyr	Pro	140	145	150
Arg	Pro	Arg	Gly	Thr	Thr	Val	Ser	Val	Gln	Gln	Leu	Phe	Ser	Thr	155	160	165
Leu	Pro	Val	Arg	His	Lys	Glu	Phe	Gln	Arg	Asn	Ile	Lys	Lys	Glu	170	175	180
Tyr	Ala	Lys	Met	Val	Gln	Val	Leu	His	Ala	Tyr	Cys	Ile	Ile	Ser	185	190	195
Ala	Gly	Ile	Arg	Val	Ser	Cys	Thr	Asn	Gln	Leu	Gly	Gln	Gly	Lys	200	205	210
Arg	Gln	Leu	Trp	Tyr	Ala	Gln	Val	Glu	Ala	Pro	Ala	Ile	Lys	Glu	215	220	225
Asn	Ile	Gly	Ser	Val	Phe	Gly	Gln	Lys	Gln	Leu	Gln	Ser	Leu	Ile	230	235	240
Pro	Phe	Val	Gln	Leu	Pro	Pro	Ser	Asp	Ser	Val	Cys	Glu	Glu	Tyr	245	250	255
Gly	Leu	Ser	Cys	Ser	Asp	Ala	Leu	His	Asn	Leu	Phe	Tyr	Ile	Ser	260	265	270
Gly	Phe	Ile	Ser	Gln	Cys	Thr	His	Gly	Val	Gly	Arg	Ser	Ser	Thr	275	280	285
Asp	Arg	Gln	Phe	Phe	Phe	Ile	Asn	Arg	Arg	Pro	Cys	Asp	Pro	Ala	290	295	300
Lys	Val	Cys	Arg	Leu	Val	Asn	Glu	Val	Tyr	His	Met	Tyr	Asn	Arg	305	310	315
His	Gln	Tyr	Pro	Phe	Val	Val	Leu	Asn	Ile	Ser	Val	Asp	Ser	Glu	320	325	330
Cys	Val	Asp	Ile	Asn	Val	Thr	Pro	Asp	Lys	Arg	Gln	Ile	Leu	Leu	335	340	345

Gln	Glu	Glu	Lys	Leu	Leu	Leu	Ala	Val	Leu	Lys	Thr	Ser	Leu	Ile
				350					355					360
Gly	Met	Phe	Asp	Ser	Asp	Val	Asn	Lys	Leu	Asn	Val	Ser	Gln	Gln
				365					370					375
Pro	Leu	Leu	Asp	Val	Glu	Gly	Asn	Leu	Ile	Lys	Met	His	Ala	Ala
				380					385					390
Asp	Leu	Glu	Lys	Pro	Met	Val	Glu	Lys	Gln	Asp	Gln	Ser	Pro	Ser
				395					400					405
Leu	Arg	Thr	Gly	Glu	Glu	Lys	Lys	Asp	Val	Ser	Ile	Ser	Arg	Leu
				410					415					420
Arg	Glu	Ala	Phe	Ser	Leu	Arg	His	Thr	Thr	Glu	Asn	Lys	Pro	His
				425					430					435
Ser	Pro	Lys	Thr	Pro	Glu	Pro	Arg	Arg	Ser	Pro	Leu	Gly	Gln	Lys
				440					445					450
Arg	Gly	Met	Leu	Ser	Ser	Ser	Thr	Ser	Gly	Ala	Ile	Ser	Asp	Lys
				455					460					465
Gly	Val	Leu	Arg	Pro	Gln	Lys	Glu	Ala	Val	Ser	Ser	Ser	His	Gly
				470					475					480
Pro	Ser	Asp	Pro	Thr	Asp	Arg	Ala	Glu	Val	Glu	Lys	Asp	Ser	Gly
				485					490					495
His	Gly	Ser	Thr	Ser	Val	Asp	Ser	Glu	Gly	Phe	Ser	Ile	Pro	Asp
				500					505					510
Thr	Gly	Ser	His	Cys	Ser	Ser	Glu	Tyr	Ala	Ala	Ser	Ser	Pro	Gly
				515					520					525
Asp	Arg	Gly	Ser	Gln	Glu	His	Val	Asp	Ser	Gln	Glu	Lys	Ala	Pro
				530					535					540
Glu	Thr	Asp	Asp	Ser	Phe	Ser	Asp	Val	Asp	Cys	His	Ser	Asn	Gln
				545					550					555
Glu	Asp	Thr	Gly	Cys	Lys	Phe	Arg	Val	Leu	Pro	Gln	Pro	Thr	Asn
				560					565					570
Leu	Ala	Thr	Pro	Asn	Thr	Lys	Arg	Phe	Lys	Lys	Glu	Glu	Ile	Leu
				575					580					585
Ser	Ser	Ser	Asp	Ile	Cys	Pro	Gln	Leu	Val	Asn	Thr	Gln	Asp	Met
				590					595					600
Ser	Ala	Ser	Gln	Val	Asp	Val	Ala	Val	Lys	Ile	Asn	Lys	Lys	Val
				605					610					615
Val	Pro	Leu	Asp	Phe	Ser	Met	Ser	Ser	Leu	Ala	Lys	Arg	Ile	Lys
				620					625					630
Gln	Leu	His	His	Glu	Ala	Gln	Gln	Ser	Glu	Gly	Glu	Gln	Asn	Tyr
				635					640					645
Arg	Lys	Phe	Arg	Ala	Lys	Ile	Cys	Pro	Gly	Glu	Asn	Gln	Ala	Ala
				650					655					660
Glu	Asp	Glu	Leu	Arg	Lys	Glu	Ile	Ser	Lys	Thr	Met	Phe	Ala	Glu
				665					670					675
Met	Glu	Ile	Ile	Gly	Gln	Phe	Asn	Leu	Gly	Phe	Ile	Ile	Thr	Thr
				680					685					690
Leu	Asn	Glu	Asp	Ile	Phe	Ile	Val	Asp	Glu	His	Ala	Thr	Asp	Glu
				695					700					705
Lys	Tyr	Asn	Phe	Glu	Met	Leu	Gln	Gln	His	Thr	Val	Leu	Gln	Gly
				710					715					720
Gln	Arg	Leu	Ile	Ala	Pro	Glu	Thr	Leu	Asn	Leu	Thr	Ala	Val	Asn
				725					730					735



Glu	Ala	Val	Leu	Ile	Glu	Asn	Leu	Glu	Ile	Phe	Arg	Lys	Asn	Gly	
				740					745					750	
Phe	Asp	Phe	Val	Ile	Asp	Glu	Asn	Ala	Pro	Val	Thr	Glu	Arg	Ala	
				755					760					765	
Lys	Leu	Ile	Ser	Leu	Pro	Thr	Ser	Lys	Asn	Trp	Thr	Phe	Gly	Pro	
				770					775					780	
Gln	Asp	Val	Asp	Glu	Leu	Ile	Phe	Met	Leu	Ser	Asp	Ser	Pro	Gly	
				785					790					795	
Val	Met	Cys	Arg	Pro	Ser	Arg	Val	Lys	Gln	Met	Phe	Ala	Ser	Arg	
				800					805					810	
Ala	Cys	Arg	Lys	Ser	Val	Met	Ile	Gly	Thr	Ala	Leu	Asn	Thr	Ser	
				815					820					825	
Glu	Met	Lys	Lys	Leu	Ile	Thr	His	Met	Gly	Glu	Met	Asp	His	Pro	
				830					835					840	
Trp	Asn	Cys	Pro	His	Gly	Arg	Pro	Thr	Met	Arg	His	Ile	Ala	Asn	
				845					850					855	
Leu	Gly	Val	Ile	Ser	Gln	Asn									
				860											

## (2) INFORMATION FOR SEQ ID NO:7:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 20 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

## (ii) MOLECULE TYPE: Oligonucleotide

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:  
GTTGAACATC TAGACGTCTC 20
- (2) INFORMATION FOR SEQ ID NO:8:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 19 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:  
TCGTGGCAGG GGTTATTCG 19
- (2) INFORMATION FOR SEQ ID NO:9:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 19 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:  
CTACCCAATG CCTCAACCG 19
- (2) INFORMATION FOR SEQ ID NO:10:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 22 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:  
GAGAACTGAT AGAAATTGGA TG 22
- (2) INFORMATION FOR SEQ ID NO:11:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 18 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

GGGACATGAG GTTCTCCG

18

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 19 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

GGGCTGTGTG AATCCTCAG

19

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 20 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

CGGTTACCA CTGTCTCGTC

20

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 18 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

TCCAGGATGC TCTCCTCG

18

(2) INFORMATION FOR SEQ ID NO:15:

- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 20 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:  
CAAGTCCTGG TAGCAAAGTC 20
- (2) INFORMATION FOR SEQ ID NO:16:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 19 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:  
ATGGCAAGGT CAAAGAGCG 19
- (2) INFORMATION FOR SEQ ID NO:17:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 22 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:  
CAACAATGTA TTCAGNAAGT CC 22
- (2) INFORMATION FOR SEQ ID NO:18:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 21 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

TTGATACAAC ACTTTGTATC G

21

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 21 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

GGAATACTAT CAGAAGGCAA G

21

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 21 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

ACAGAGCAAG TTACTCAGAT G

21

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 20 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

GTACACAATG CAGGCATTAG

20

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 21 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

AATGTGGATG TTAATGTGCA C

21

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 19 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

CTGACCTCGT CTTCCTAC

19

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 19 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

CAGCAAGATG AGGAGATGC

19

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 21 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

GGAAATGGTG GAAGATGATT C

21

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 16BASE PAIRS

- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

CTTCTCAACA CCAAGC

16

(2) INFORMATION FOR SEQ ID NO:27:

(i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 21 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

GAAATTGATG AGGAAGGGAA C

21

(2) INFORMATION FOR SEQ ID NO:28:

(i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 22 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

CTTCTGATTG ACAACTATGT GC

22

(2) INFORMATION FOR SEQ ID NO:29:

(i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 22 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:29:

CACAGAAGAT GGAAATATCC TG

22

## (2) INFORMATION FOR SEQ ID NO:30:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 20 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

GTGTTGCTAG CACTTAAGAC

20

## (2) INFORMATION FOR SEQ ID NO:31:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 20 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

TTTCCCATAT TCTTCACTTG

20

## (2) INFORMATION FOR SEQ ID NO:32:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 19 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:32:

GTAACATGAG CCACATGGC

19

## (2) INFORMATION FOR SEQ ID NO:33:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 19 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide



- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:33:  
CCACTGTCTC GTCCAGCCG 19
- (2) INFORMATION FOR SEQ ID NO:34:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 26 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:34:  
CGGGATCCAT GTCGTTCTG GCAGGG 26
- (2) INFORMATION FOR SEQ ID NO:35:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 26 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:35:  
GCTCTAGATT AACACCTCTC AAAGAC 26
- (2) INFORMATION FOR SEQ ID NO:36:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 21 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:36:  
GCATCTAGAC GTTTCCTTGG C 21
- (2) INFORMATION FOR SEQ ID NO:37:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 20 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:37:

CATCCAAGCT TCTGTTCCCG

20

(2) INFORMATION FOR SEQ ID NO:38:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 19 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:38:

GGGGTGCAGC AGCACATCG

19

(2) INFORMATION FOR SEQ ID NO:39:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 20 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:39:

GGAGGCAGAA TGTGTGAGCG

20

(2) INFORMATION FOR SEQ ID NO:40:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 19 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:40:

TCCCAAAGAA GGACTTGCT

19

(2) INFORMATION FOR SEQ ID NO:41:

- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 22 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:41:  
AGTATAAGTC TTAAGTGCTA CC 22
- (2) INFORMATION FOR SEQ ID NO:42:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 20 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:41:  
TTTATGGTTT CTCACCTGCC 20
- (2) INFORMATION FOR SEQ ID NO:43:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 19 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:43:  
GTTATCTGCC CACCTCAGC 19
- (2) INFORMATION FOR SEQ ID NO:44:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 59 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:44:

GGATCCTAAT ACGACTCACT ATAGGGAGAC CACCATGGCA TCTAGACGTT TCCCTTGGC

59

## (2) INFORMATION FOR SEQ ID NO:45:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 20 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

## (ii) MOLECULE TYPE: Oligonucleotide

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:45:

CATCCAAGCT TCTGTTCCCG

20

## (2) INFORMATION FOR SEQ ID NO:46:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 56 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

## (ii) MOLECULE TYPE: Oligonucleotide

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:46:

GGATCCTAAT ACGACTCACT ATAGGGAGAC CACCATGGGG GTGCAGCAGC ACATCG

56

## (2) INFORMATION FOR SEQ ID NO:47:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 20 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

## (ii) MOLECULE TYPE: Oligonucleotide

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:47:

GGAGGCAGAA TGTGTGAGCG

20

## (2) INFORMATION FOR SEQ ID NO:48:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 28 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:48:  
CGGGATCCAT GAAACAATTG CCTGCGGC 28
- (2) INFORMATION FOR SEQ ID NO:49:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 26 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:49:  
GCTCTAGACC AGACTCATGC TGTTTT 26
- (2) INFORMATION FOR SEQ ID NO:50:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 26 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:50:  
CGGGATCCAT GGAGCGAGCT GAGAGC 26
- (2) INFORMATION FOR SEQ ID NO:51:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 23 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:51:  
GCTCTAGAGT GAAGACTCTG TCT 23
- (2) INFORMATION FOR SEQ ID NO:52:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 20 BASE PAIRS

- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:52:

AAGCTGCTCT GTTAAAAGCG

20

(2) INFORMATION FOR SEQ ID NO:53:

(i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 18 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:53:

GCACCAGCAT CCAAGGAG

18

(2) INFORMATION FOR SEQ ID NO:54:

(i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 19 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:54:

CAACCATGAG ACACATCGC

19

(2) INFORMATION FOR SEQ ID NO:55:

(i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 20 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:55:

AGGTTAGTGA AGACTCTGTC

20

## (2) INFORMATION FOR SEQ ID NO:56:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 53 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:56:

GGATCCTAAT ACGACTCACT ATAGGGAGAC CACCATGGAA CAATGCCTG CGG

53

## (2) INFORMATION FOR SEQ ID NO:57:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 18 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:57:

CCTGCTCCAC TCATCTGC

18

## (2) INFORMATION FOR SEQ ID NO:58:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 60 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:58:

GGATCCTAAT ACGACTCACT ATAGGGAGAC CACCATGGAA GATATCTTAA AGTTAATCCG

60

## (2) INFORMATION FOR SEQ ID NO:59:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 21 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:59:

GGCTTCTTCT ACTCTATATG G

21

(2) INFORMATION FOR SEQ ID NO:60:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 58 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:60:

GGATCCTAAT ACGACTCACT ATAGGGAGAC CACCATGGCA GGTCTTGAAA ACTCTTCG

58

(2) INFORMATION FOR SEQ ID NO:61:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 21 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:61:

AAAACAAGTC AGTGAATCCT C

21

(2) INFORMATION FOR SEQ ID NO:62:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 20 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:62:

AAGCATCT GTTTCTGCTG

20

(2) INFORMATION FOR SEQ ID NO:63:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 20 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE



(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:63:

ACGAGTAGAT TCCTTTAGGC

20

(2) INFORMATION FOR SEQ ID NO:64:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 19 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:64:

CAGAACTGAC ATGAGAGCC

19

(2) INFORMATION FOR SEQ ID NO:65:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 52 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:65:

GGATCCTAAT ACGACTCACT ATAGGGAGAC CACCATGGAG CGAGCTGAGA GC

52

(2) INFORMATION FOR SEQ ID NO:66:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 20 BASE PAIRS

(B) TYPE: NUCLEIC ACID

(C) STRANDEDNESS: SINGLE

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:66:

AGGTTAGTGA AGACTCTGTC

20

(2) INFORMATION FOR SEQ ID NO:67:

- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 17 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:67:  
CTGAGGTCTC AGCAGGC 17
- (2) INFORMATION FOR SEQ ID NO:68:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 57 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:68:  
GGATCCTAAT ACGACTCACT ATAGGGAGAC CACCATGGTG TCCATTTCCTCA GACTGCG 57
- (2) INFORMATION FOR SEQ ID NO:69:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 20 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:69:  
AGGTTAGTGA AGACTCTGTC 20
- (2) INFORMATION FOR SEQ ID NO:70:
- (i) SEQUENCE CHARACTERISTICS  
(A) LENGTH: 21 BASE PAIRS  
(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:70:

TTATTTGGCA GAAAAGCAGA G

21

## (2) INFORMATION FOR SEQ ID NO:71:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 21 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:71:

TTAAAAGACT AACCTCTTGC C

21

## (2) INFORMATION FOR SEQ ID NO:72:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 21 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:72:

CTGCTGTTAT GAACAATATG G

21

## (2) INFORMATION FOR SEQ ID NO:73:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 19 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:73:

CAGAAGCAGT TGCAAAGCC

19

## (2) INFORMATION FOR SEQ ID NO:74:

## (i) SEQUENCE CHARACTERISTICS

- (A) LENGTH: 20 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:74:
- AAACCGTACT CTTACACAC 20
- (2) INFORMATION FOR SEQ ID NO:75:
- (i) SEQUENCE CHARACTERISTICS
- (A) LENGTH: 20 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:75:
- GAGGAAAAGC TTTTGTTGGC 20
- (2) INFORMATION FOR SEQ ID NO:76:
- (i) SEQUENCE CHARACTERISTICS
- (A) LENGTH: 18 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:76:
- CAGTGGCTGC TGACTGAC 18
- (2) INFORMATION FOR SEQ ID NO:77:
- (i) SEQUENCE CHARACTERISTICS
- (A) LENGTH: 19 BASE PAIRS
- (B) TYPE: NUCLEIC ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR
- (ii) MOLECULE TYPE: Oligonucleotide
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:77:
- TCCAGAACCA AGAAGGAGC 19
- (2) INFORMATION FOR SEQ ID NO:78:
- (i) SEQUENCE CHARACTERISTICS
- (A) LENGTH: 16 BASE PAIRS

(B) TYPE: NUCLEIC ACID  
(C) STRANDEDNESS: SINGLE  
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:78:

TGAGGTCTCA GCAGGC

16

WHAT IS CLAIMED IS:

1. An isolated polynucleotide selected from the group consisting of:
  - (a) a polynucleotide encoding a polypeptide having the deduced amino acid sequence of SEQ ID No. 2 or a fragment, analog or derivative of said polypeptide;
  - (b) a polynucleotide encoding a polypeptide having the amino acid sequence encoded by the cDNA contained in ATCC Deposit No. 75649;
  - (c) a polynucleotide encoding a polypeptide having the deduced amino acid sequence of SEQ ID No. 4 or a fragment, analog or derivative of said polypeptide;
  - (d) a polynucleotide encoding a polypeptide having the amino acid sequence encoded by the cDNA contained in ATCC Deposit No. 75651;
  - (e) a polynucleotide encoding a polypeptide having the deduced amino acid sequence of SEQ ID No. 6 or a fragment, analog or derivative of said polypeptide; and
  - (f) a polynucleotide encoding a polypeptide having the amino acid sequence encoded by the cDNA contained in ATCC Deposit No. 75650.
2. The polynucleotide of Claim 1 wherein the polynucleotide is DNA.
3. The polynucleotide of Claim 1 wherein the polynucleotide is RNA.
4. The polynucleotide of Claim 1 wherein the polynucleotide is genomic DNA.
5. The polynucleotide sequence of claim 1 for use in analyzing a sample for mutation of a polynucleotide sequence encoding a human mismatch repair protein comprising:

a polynucleotide sequence of at least 15 and no more than 30 consecutive bases of the polynucleotide sequence of ATTC Deposit No. 75649.

6. The polynucleotide sequence of claim 1 for use in analyzing a sample for mutation of a polynucleotide sequence encoding a human mismatch repair protein comprising:

a polynucleotide sequence of at least 15 and no more than 30 consecutive bases of the the polynucleotide sequence of ATTC Deposit No. 75651.

7. The polynucleotide sequence of claim 1 for use in analyzing a sample for mutation of a polynucleotide sequence encoding a human mismatch repair protein comprising:

a polynucleotide sequence of at least 15 and no more than 30 consecutive bases of the the polynucleotide sequence of ATTC Deposit No. 75650.

8. The polynucleotide of Claim 2 wherein said polynucleotide encodes a polypeptide having the deduced amino acid sequence of SEQ ID No. 2.

9. The polynucleotide of Claim 2 wherein said polynucleotide encodes a polypeptide having the deduced amino acid sequence of SEQ ID No. 4.

10. The polynucleotide of Claim 2 wherein said polynucleotide encodes a polypeptide having the deduced amino acid sequence of SEQ ID No. 6.

11. The polynucleotide of Claim 2 wherein said polynucleotide encodes a polypeptide encoded by the cDNA of ATCC Deposit No. 75649.

12. The polynucleotide of Claim 2 wherein said polynucleotide encodes a polypeptide encoded by the cDNA of ATCC Deposit No. 75651.
13. The polynucleotide of Claim 2 wherein said polynucleotide encodes a polypeptide encoded by the cDNA of ATCC Deposit No. 75650.
14. The polynucleotide of Claim 1 having the coding sequence of SEQ ID No. 1.
15. The polynucleotide of Claim 1 having the coding sequence of SEQ ID No. 3.
16. The polynucleotide of Claim 1 having the coding sequence of SEQ ID No. 5).
17. A vector containing the DNA of Claim 2.
18. A host cell genetically engineered with the vector of Claim 17.
19. A process for producing a polypeptide comprising: expressing from the host cell of Claim 18 the polypeptide encoded by said DNA.
20. A process for producing cells capable of expressing a polypeptide comprising genetically engineering cells with the vector of Claim 17.
21. An isolated DNA hybridizable to the DNA of Claim 2 and encoding a polypeptide having hMLH1 activity.
22. An isolated DNA hybridizable to the DNA of Claim 2 and encoding a polypeptide having hMLH2 activity.



23. An isolated DNA hybridizable to the DNA of Claim 2 and encoding a polypeptide having hMLH3 activity.

24. A polypeptide selected from the group consisting of:

(a) a polypeptide having the deduced amino acid sequence of SEQ ID No. 2 and fragments, analogs and derivatives thereof;

(b) a polypeptide encoded by the cDNA of ATCC Deposit No. 75649 and fragments, analogs and derivatives of said polypeptide;

(c) a polypeptide having the deduced amino acid sequence of SEQ ID No. 4 and fragments, analogs and derivatives thereof;

(d) a polypeptide encoded by the cDNA of ATCC Deposit No. 75651 and fragments, analogs and derivatives of said polypeptide;

(e) a polypeptide having the deduced amino acid sequence of SEQ ID No. 6 and fragments, analogs and derivatives thereof; and

(f) a polypeptide encoded by the cDNA of ATCC Deposit No. 75650 and fragments, analogs and derivatives of said polypeptide.

25. The polypeptide of Claim 15 wherein the polypeptide is hMLH1 having the deduced amino acid sequence of SEQ ID No. 2.

26. The polypeptide of Claim 14 wherein the polypeptide is hMLH2 having the deduced amino acid sequence of SEQ ID No. 4.

27. The polypeptide of Claim 14 wherein the polypeptide is hMLH3 having the deduced amino acid sequence of SEQ ID No. 6.

28. A process for diagnosing a susceptibility to cancer comprising:

determining from a sample derived from a human patient a mutation in a human mismatch repair gene, said human mismatch repair gene comprising the polynucleotide sequence of claim 8.

29. A process for diagnosing a susceptibility to cancer comprising:

determining from a sample derived from a human patient a mutation in a human mismatch repair gene, said human mismatch repair gene comprising the DNA of claim 9.

30. A process for diagnosing a susceptibility to cancer comprising:

determining from a sample derived from a human patient a mutation in a human mismatch repair gene, said human mismatch repair gene comprising the DNA of claim 10.

31. A process for diagnosing a susceptibility to cancer comprising:

determining from a sample derived from a human patient a mutation in a human DNA mismatch repair gene which encodes the human homolog of a bacterial mutL DNA mismatch repair gene.

-40                      -20                      1    FIG. 1A

```

gttgaacatctagacgtttccttggctcttctggcgccaaaATGTCGTCGTGGCAGGGG
+-----+-----+-----+-----+-----+-----+-----+-----+
caacttgtagatctgcaaaaggaaccgagaacgcgcggtttTACAGCAAGCACCGTCCCC
                                     M S F V A G V
20                               40                               60

TTATTCCGGCGCTGGACGAGACAGTGGTGAACCGCATCGGGGGGGAAGTTATCCAGC
+-----+-----+-----+-----+-----+-----+-----+-----+
AATAAGCCGCGCACCTGCTCTGTCAACCACTTGGCGTAGCGCGCCGCCCTTCAATAGGTCG
I R R L D E T V V N R I A A G E V I Q R
80                               100                               120

GGCCAGCTAATGCTATCAAAGAGATGATTGAGAACTGTTAGATGCAAAATCCACAAGTA
+-----+-----+-----+-----+-----+-----+-----+-----+
CCGGTCGATTACGATAGTTTCTCTACTAATCTTGACAAATCTACGTTTTAGGTGTTTCAT
P A N A I K E M I E N C L D A K S T S I
140                               160                               180

TTCAAGTGATTGTTAAAGAGGGAGGCCCTGAAGTTGATTTCAGATCCAGACAATGGCACCG
+-----+-----+-----+-----+-----+-----+-----+-----+
AAGTTCACTAACAAATTCTCCCTCCGGACTTCAACTAAGTCTAGGTTCTGTTACCGTGCC
Q V I V K E G G L K L I Q I Q I Q D N G T G
200                               220                               240

. MATCH WITH FIG. 1B

```

MATCH WITH FIG. 1A

FIG. 1B

```

GGATCAGGAAAGATCTGGATATTGTATGTGAAAGGTTCACTACTAGTAAACTGCAGT
+-----+-----+-----+-----+-----+-----+-----+-----+
CCTAGTCCTTTCTCTAGACCTATAACATACACTTCCCAAGTGATGATCATTTGACGTCA
I R K E D L D I V C E R F T T S K L Q S
260                               280                               300

CCTTTGAGGATTAGCCAGTATTTCTACCTATGGCTTTCGAGGTGAGGCTTTGGCCAGCA
+-----+-----+-----+-----+-----+-----+-----+-----+
GGAACCTCCTAAATCGGTCAATAAAGATGGATACCGAAGCTCCACTCCGAAACCGGTCGT
F E D L A G I S T Y G F R G E A L A S I
320                               340                               360

TAAGCCATGTGGCTCATGTTACTATTACAAACGAAACAGCTGATGGAAAGTGTGCATACA
+-----+-----+-----+-----+-----+-----+-----+-----+
ATTCCGGTACACCGAGTACAATGATAATGTTGCTTTTGTGCGACTACCTTTCACACGTATGT
S H V A H V T I T T K T A D G K C A Y R
380                               400                               420

GAGCAAGTTACTCAGATGGAAAACTGAAAGCCCCCTCCTAAACCATGTGCTGGCAATCAAG
+-----+-----+-----+-----+-----+-----+-----+-----+
CTCGTTCAATGAGTCTACCTTTTGACTTTTCGGGGAGGATTGTTGTTACACGACCGTTAGTTC
A S Y S D G K L K A P P K P C A G N Q G
440                               460                               480

GGACCCAGATCACGGTGGAGGACCTTTTTTTACAACATAGCCACGAGGAGAAAGCTTTAA
+-----+-----+-----+-----+-----+-----+-----+-----+

```

MATCH WITH FIG. 1C

MATCH WITH FIG. 1B

FIG. 1C

CCTGGGTCCTAGTGCCACCTCCTGGAAAAATGTTGTATCGGTGCTCCTCTTTCGAAATT  
 T Q I T V E D L F Y N I A T R R K A L K  
 500 520 540

AAATCCAAGTGAAGAAATATGGGAAAAATTTGGAAATTGTTGCCAGGTATTCAGTACACA  
 +-----+-----+-----+-----+-----+-----+-----+-----+  
 TTTTAGGTTCACTTCTTATACCCCTTTTAAACCTTCAACAACCGTTCATAAGTCATGTGT  
 N P S E E Y G K I L E V V G R Y S V H N  
 560 580 600

ATGCAGGCATTAGTTTCTCAGTTAAAAACAAGGAGAGACAGTAGCTGATGTTAGGACAC  
 +-----+-----+-----+-----+-----+-----+-----+-----+  
 TACGTCGGTAATCAAAGAGTCAATTTTGTTCCTCTCTGTGTCATCGACTACAAATCCCTGTG  
 A G I S F S V K K Q G E T V A D V R T L  
 620 640 660

TACCCAATGCCCTCAACCGTGGACAATATTCGCTCCGTCCTTTGGAAATGCTGTAGTCGAG  
 +-----+-----+-----+-----+-----+-----+-----+-----+  
 ATGGGTTACGGAGTTGGCACCTGTTATAAGCGAGGCAGAAACCCCTTACGACAATCACCTC  
 P N A S T V D N I R S V F G N A V S R E  
 680 700 720

AACTGATAGAAATTGGATGTGAGGATAAAACCCCTAGCCCTTCAAAATGAATGGTTACATAT  
 +-----+-----+-----+-----+-----+-----+-----+-----+  
 TTGACTATCTTTAACCTACACTCCTATTTTGGGATCGGAAGTTTACTTACCAATGTATA  
 L I E I G C E D K T L A F K M N G Y I S  
 MATCH WITH FIG. 1D

## FIG. 1D

MATCH WITH FIG. 1C  
 740  
 CCAATGCAAACTACTCAGTGAAGAAGTGCACTCTTCTTACTCTTCAATCAACCATCGTCTGG  
 +-----+-----+-----+-----+-----+-----+-----+-----+  
 GGTACGTTGATGAGTCACTTCTTCACGTAGAAGAAATGAGAAGTAGTTGGTAGCAGACC  
 N A N Y S V K K C I F L L F I N H R L V  
 800 820 840  
 TAGAATCAAACTTCCTTGAGAAAAGCCATAGAAACAGTGATGCAGCCTATTGCCCCAAA  
 +-----+-----+-----+-----+-----+-----+-----+-----+  
 ATCTTAGTTGAAGGAACTCTTTTCGGTATCTTGTACATACGTCGGATAAACGGGTTT  
 E S T S L R K A I E T V Y A A Y L P K N  
 860 880 900  
 ACACACACCCATTCCCTGTACCTCAGTTTAGAAATCAGTCCCAGAAATGTGGATGTTAATG  
 +-----+-----+-----+-----+-----+-----+-----+-----+  
 TGTGTGGGTAAAGGACATGGAGTCAAAATCTTTAGTCAGGGGTCTTACACCTACAATTAC  
 T H P P L Y L S L E I S P Q N V D V N V  
 920 940 960  
 TGCACCCACAAAGCATGAAGTTCACTTCCCTGCACGAGGAGAGCATCCTGGAGCGGGTGC  
 +-----+-----+-----+-----+-----+-----+-----+-----+  
 ACGTGGGGTGTTCGTACTTCAAGTGAAGGACGTGCTCCTCTCGTAGGACCTCGCCACG  
 H P T K H E V H F L H E E S I L E R V Q  
 980 1000 1020  
 MATCH WITH FIG. 1E

## FIG. 1E

MATCH WITH FIG. 1D

```

AGCAGCACATCGAGAGCAAGCTCCTGGGCTCCAATTCCTCCAGGATGTAATCACCACCCAGA
+-----+-----+-----+-----+-----+-----+-----+-----+
TCGTCGTGTAGCTCTCGTTCGAGGACCCGAGGTTAAGGAGGTCCTACATGAAGTGGGTCT
Q H I E S K L L L G S N S S R M Y F T Q T
1040                                     1060                               1080

CTTTGCTACCAAGGACTTGCTGGCCCCCTCTGGGGAGATGGTTAAATCCACAAGTCTCA
+-----+-----+-----+-----+-----+-----+-----+-----+
GAAACGATGGTCTGAACGACGGGGGAGACCCCTCTACCAATTAGGTGTTGTTTCAGACT
L L H G L A A P S G E M V K S T T S L T
1100                                     1120                               1140

CCTCGTCTTCTACTTCTGGAAAGTAGTGATAAGGTCTATGCCACCACAGATGGTTCGTACAG
+-----+-----+-----+-----+-----+-----+-----+-----+
GGAGCAGAAAGATGAAGACCTTCATCATACTATCCAGATACGGGTGGTCTACCAAGCATGTC
S S S T S G S S D K V Y A H Q M V R T D
1160                                     1180                               1200

ATCCCCGGGAACAGAAGCTTGATGCATTTCTGCAGCCTCTGAGCAAAACCCCTGTCCAGTC
+-----+-----+-----+-----+-----+-----+-----+-----+
TAAGGGCCCTTGTCTTCGAACTACGTAAAGACGTCCGAGACTCGTTTGGGGACAGGTCAG
S R E Q Q K I D A F L Q P L S K P L S S Q
1280                                     1240                               1260

AGCCCCAGGCCATTGTCACAGAGGATAAGACAGATATTTCTAGTGCCAGGGCTAGGCAGC
MATCH WITH FIG. 1F

```

## FIG. 1F

MATCH WITH FIG. 1E

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+
TCGGGTCCGGTAACAGTGTCTCCTATTCTGTCTATAAAGATCACCGTCCCGATCCGTCG
P Q A I V T E D K T D I S S G R A R Q Q
1280                               1300                               1320

.
AAGATGAGGAGATGCTTGAACTCCCAGCCCCCTGCTGAAGTGGCTGCCAAAATCAGAGCT
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
TTCTACTCCTCTACGAACCTTGAGGGTCGGGACGACTTCACCGACGGTTTTAGTCTCGA
D E E M L E L P A P A E V A A K N Q S L
1340                               1360                               1380

.
TGGAGGGGGATACAACAAAGGGGACTTCAGAAATGTCAGAGAAGAGAGACCTACTTCCA
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
ACCTCCCCCTATGTTGTTTCCCTGAAGTCTTTACAGTCTCTCTCTCCTGGATGAAGGT
E G D T T K G T S E M S E K R G P T S S
1400                               1420                               1440

.
GCAACCCAGAAAGAGACATCGGGAAGATTCTGATCTCCAATCCTCGAAGATGATCCC
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
CGTTGGGGTCTTTCTCTGTAGCCCTTCTAAGACTACACCTTTACCACCTTCTACTAAGGG
N P R K R H R E D S D V E M V E D D S R
1460                               1480                               1500

.
GAAAGGAAATGACTGCAGCTTGTAACCCCCGGGAGAGGATCATTAACCTCAGTGTTT
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
CTTTCCTTACTGACGTCGAACATGGGGGCCCTCTCCTAGTAATGGAGTGATCACAAA

```

MATCH WITH FIG. 1G



## FIG. 1G

MATCH WITH FIG. 1F

```

K E M T A A C T P R R R I I N L T S V L
1520                                     1560
TGAGTCTCCAGGAAGAAATTAATGAGCAGGGACATGAGGTTCTCCGGGAGATGTTGCATA
+-----+-----+-----+-----+-----+-----+-----+-----+
ACTCAGAGGTCCTTCTTAATTAATCTGTCCTCGTACTCCAGAGAGGCCCTCTACAACGTAT
S L Q E E I N E Q G H E V L R E M L H N
1580                                     1620
ACCACTCCTTCGTGGGCTGTGTGAATCCTCAGTGGGCCCTTGGCACAGCATCAAACCAAGT
+-----+-----+-----+-----+-----+-----+-----+-----+
TGGTGAGGAAGCACCACACACTTAGGAGTCACCCCGGAACCGTGTGCTAGTTGCTTCA
H S F V G C V N P Q W A L A Q H Q T K L
1640                                     1680
TATACCTTCTCAACACCACCAAGCTTAGTGAAGAACTGTTCTACCAGATACTCATTTATG
+-----+-----+-----+-----+-----+-----+-----+-----+
ATATGGAAGAGTTGTGGTGGTTTCGAATCAGTCTTGACAAAGATGGTCTATGAGTAAATAC
Y L, L N T T K L S E E L F Y Q I L I Y D
1700                                     1740
ATTTTGCCCAATTTTGGTGTCTCAGGTTATCGGAGCCAGCACCGCTCTTGACCTTGCCA
+-----+-----+-----+-----+-----+-----+-----+-----+
TAAACGGTTAAACACCAAGAGTCCAATAGCCTCGGTGCGGAGAACTGGAACGGT
F A N F G V L R L S E P A P L F D L A M
1760                                     1800

```

MATCH WITH FIG. 1H

7/41

## FIG. 1H

MATCH WITH FIG. 1G

```

TGCTTCCCTTAGATAGTCCAGAGAGTGGCTGGACAGAGGAAGATGGTCCCAAGAAGGAC
+-----+-----+-----+-----+-----+-----+-----+-----+
ACGAACGGAATCTATCAGGTCCTCACCACCTGTCTCCTTCTACCAAGGTTTCTTCCCTG
L A L D S P E S G W T E E D G P K E G L
1820                                     1840 1860

TTGCTGAATACATTGTTGAGTTTCTGAAGAAGAGGCTGAGATGCTTGCAGACTATTCT
+-----+-----+-----+-----+-----+-----+-----+-----+
AACGACTTATGTAACAACCTCAAGACTTCTTCTCCGACTCTACGAACGCTCTGATAAAGA
A E Y I V E F L K K A E M L A D Y F S
1880                                     1900 1920

CTTTGGAAATTGATGAGGAAGGGAACCTGATTGGATTACCCCTTCTGATTGACAACTATG
+-----+-----+-----+-----+-----+-----+-----+-----+
GAAACCTTAACTACTCCTTCCCTTGGACTAACCTAATGGGGAAGACTAACTGTTGATAC
L E I D E E G N L I G L P L L T D N Y V
1940                                     1960 1980

TGCCCCCTTTGGAGGGACTCCCTATCTTCAATTCTTCCACTAGCCACTGAGGTGAATTGGG
+-----+-----+-----+-----+-----+-----+-----+-----+
ACGGGGAAACCTCCCTGACGGATAGAAGTAAGAAGCTGATCGGTGACTCCACTTAACCC
P P L E G L P I F I L R L A T E V N W D
2000                                     2020 2040

ACGAAGAAAGGAATGTTTGAAGCCCTCAGTAAAGAATGCGCTATGTTCTATTCATCC
MATCH WITH FIG. 1I

```

## FIG. II

MATCH WITH FIG. 1H

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
TGCTTCTTTCCCTTACAAAACCTCTCGGAGTCATTCTTACGGGATACAAGATAAGGTAGG
E E K E C F E S L S K E C A M F Y S I R
2060                                2080                                2100

.
GGAAGCAGTACATATCTGAGGAGTCGACCCCTCTCAGGCCAGAGTGAAAGTGCCTGGCT
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
CCTTCGTCATGTATAGACTCCTCAGCTGGGAGAGTCCGGTCGTCTCACTTCACGGACCGA
K Q Y I S E E S T L S G Q Q S E V P G S
2120                                2140                                2160

.
CCATTCCAAAACCTCTGGAAGTGGACTGTGGAACACATTGTCTATAAAGCCTTGCGCTCAC
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
GGTAAGGTTTGAGGACCTTACCTGACACCTTGTGTAACAGATATTTCGGAACGCGAGTG
I P N S W K W T V E H I V Y K A L R S H
2180                                2200                                2220

.
ACATTCTGCCCTCCTAAACATTTCACAGAAGATGGAAATATCCTGCAGCTTGCTAACCTGC
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
TGTAAGACGGAGGATTTGTAAAGTGCTTCTACCTTTATAGGACGTCGAACGATTGGACG
I L P P K H S T E D G N I L Q L A N L P
2240                                2260                                2280

.
CTGATCTATACAAAGTCTTTGAGAGGTGTTAAATATGGTTATTATGCACGTGGGATGT
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
GACTAGATATGTTTCAGAAACCTCTCCACAATTTATACCAATAAATACGTGACACCCCTACA
MATCH WITH IG. 1J

```

רפ"א

MATCH WITH FIG. 11

D L Y K V F E R C \*  
2300 2320 2340  
GTTCTTCTTCTCTGTATCCGATACAAAGTGTGTATCAAAAGTGTGATATACAAAGTGT  
+-----+-----+-----+-----+-----+-----+-----+-----+  
CAAGAAGAAAGAGACATAAGGCTATGTTTCACAACATAGTTTCACACTATATGTTTCACA  
2360 2380 2400  
ACCAACATAAGTGTGGTAGCACTTAAGACTTATACTTGCCCTTCTGACAGTATTCCTTTA  
+-----+-----+-----+-----+-----+-----+-----+-----+  
TGGTTGTATTTCACAACCATCGTGAATTCTGAAATATGAACGGAAGACTATCATAAAGGAAAT  
2420 2440 2460  
TACACATGTGATTGATTATAAATAAATAGATGTGTCTTAACATAAAAAAATAAAAAA  
+-----+-----+-----+-----+-----+-----+-----+-----+  
ATGTGTCACCTAACTAATATTATTATCTACACAGAAATTGTATTTTTTT  
2480  
AAAAA  
+-----  
TTTTTT

Polynucleotide and deduced amino acid sequence of hMLH3: **FIG. 2A**

```

-70      -50      -30
GGCAGAGTGGCTGCTTGGCGGTAGTGATGGTAATTGCCCTGCCCTCGCGCTAGCAGCAAG
-----+-----+-----+-----+-----+-----+-----+
CCGTGCTCACCAGACGACCGCGATCACCTACCATTAACGGACGGAGCGGATCGTCGTTTC
-10      10      30
CTGCTCTGTAAAGCGAAATGAACAATTGCCCTGCGGCAACAGTTCGACTCCTTTCAA
-----+-----+-----+-----+-----+-----+-----+
GACGAGACAATTTTCGCTTTTACTTTGTAAACGGACGCCGTTGTCAAGCTGAGGAAAGTT
                    M K Q L P A A T V R L L S S
                    50      70      90
GTTCTCAGATCATCACTTCGGTGGTCAGTGTGTAAAGAGCTTATTGAAAACTCCTTGG
-----+-----+-----+-----+-----+-----+-----+
CAAGAGCTAGTAGTGAAGCCACCAGTCACAACATTTTCTCGAATAACTTTTGAGGAACC
S Q I I T S V V S V V K E L I E N S L D
110      130      150
ATGCTGGTGCCACAAGCGTAGATGTAAACTGGAGAACTATGGATTGTATAAAATTGAGG
-----+-----+-----+-----+-----+-----+-----+
TAGACCAACGGTGTTCGCATCTACAATTTGACCTCTTGATACCTAAACTATTTTAACTCC
A G A T S V D V K L E N Y G F D K I E V
170      190      210

```

MATCH WITH FIG. 2B

11/41

SUBSTITUTE SHEET (RULE 26)

MATCH WITH FIG. 2A                      FIG. 2B

```

TGCAGATAACGGGGAGGTATCAAGGCTGTGATGCACCTGTAATGGCAATGAAGTACT
-----+-----+-----+-----+-----+-----+-----+
ACGCTCTATTGCCCTCCCATAGTTCGACAACTACGTGGACATTACCGTTACTTCATGA
R D N G E G I K A V D A P V M A M K Y Y
230                               250 270

ACACCTCAAAATAATAGTCATGAAGATCTTGAAAATTGACAACTTACGGTTTTCGTG
-----+-----+-----+-----+-----+-----+-----+
TGTGGAGTTTATTATCAGTACTTCTAGAACTTTTAAACTGTTGAATGCCAAAGCAC
T S K I N S H E D L E N L T T Y G F R G
290                               310 330

GAGAAGCCTTGGGGTCAATTGTGTATAGCTGAGGTTTAAATTACAACAAGACGGCTG
-----+-----+-----+-----+-----+-----+-----+
CTCTTCGGAACCCAGTTAAACAACATATCGACTCCAAATTAATGTTGTTCTTGCCGAC
E A L G S I C C I A E V L I T T R T A A
350                               370 390

CTGATAATTTAGCACCCAGTATGTTTATAGTGCAGTGGCCACATACTTTCTCAGAAAC
-----+-----+-----+-----+-----+-----+-----+
GACTATTAAATCGTGGTCATACAAAATCTACCGTCCCGGTGTATGAAAGAGTCTTGG
D N F S T Q Y V L D G S G H I L S Q K P
MATCH WITH FIG. 2C

```

12/41

SUBSTITUTE SHEET (RULE 26)

13 / 41

## FIG. 2D

MATCH WITH FIG. 2C

A V I W Q K S R V S D H K M A L M S V L  
 650 670 690

TGGGACTGCTGTATGAACAATATGAATCCTTTCAGTACCACCTCTGAAGAATCTCAGA  
 -----+-----+-----+-----+-----+  
 ACCCTGACGACAATACTTGTTATACCTTAGGAAAGTCATGGTGAGACTTCTTAGAGTCT  
 G T A V M N N M E S F Q Y H S E S Q I  
 710 730 750

TTTATCTCAGTGGATTTCCTCCAAAGTGTGATGCAGACCACCTCTTTCAGTCTTTCAA  
 -----+-----+-----+-----+-----+  
 AAATAGAGTCACCTAAAGAAGTTTCACACTACGCTCGTGAGAAAGTCAGATCAGAAAGTT  
 Y L S G F L P K C D A D H S F T S L S T  
 770 790 810

CACCAGAAAGATTTCATCTTCATAAACAGTCGACCAGTACATCAAAAAGATATCTTAA  
 -----+-----+-----+-----+-----+  
 GTGGTCTTCTTCAAGTAGAAGTATTGTGCAGCTGGTCATGTAGTTTCTTATAGAATT  
 P E R S F I F I N S R P V H Q K D I L K  
 830 850 870

AGTTAATCCGACATCATTAACAATCTGAAATGCCCTAAAGGAATCTACTCGTTGTATCCCTG  
 -----+-----+-----+-----+-----+  
 MATCH WITH FIG. 2E



## FIG. 2E

MATCH WITH FIG. 2D

TC AATTAGGCTGTAGTAATGTTAGACTTTACGGATTTCCTTAGATGAGCAACATAGGAC  
 L I R H H Y N L K C L K E S T R L Y P V  
 890 910 930

TTTCTTCTGAAAATCGATGTTCCCTACAGCTGATGTTGATGTAAATTTAACACCAGATA  
 -----+-----+-----+-----+-----+-----+-----+  
 AAAAGAAAGACTTTTAGCTACAAGGATGTCGACTACAACTACATTTAAATTTGTGGTCTAT  
 F F L K I D V P T A D V D V N L T P D K  
 950 970 990

AAAGCCAAGTATTATTACAAAATAAGGAATCTGTTTAAATGCTCTTGAAATCTGATGA  
 -----+-----+-----+-----+-----+-----+-----+  
 TTTCGGTTCATAATAATGTTTATTTCCTTAGACAAAATAACGAGAACTTTTAGACTACT  
 S Q V L L Q N K E S V L I A L E N L M T  
 1010 1030 1050

CGACTTGTTATGGACCATTACCTAGTACAAATTCTTATGAAAATAATAAACAGATGTTT  
 -----+-----+-----+-----+-----+-----+-----+  
 GCTGAACAATACCTGGTAATGGATCATGTTTAAGAACTTTTATTATTGTCTACAAA  
 T C Y G P L P S T N S Y E N N K T D V S  
 1070 1090 1110

· MATCH WITH FIG. 2F ·

**MATCH WITH FIG. 2E**

**THE**

A A D I V L S K T A E T D V I. F N K V F

1130  
1150  
1170

[illegible]

S S G K N Y S N V D T S V T P E C N P V

1190 1191 1192 1193 1194 1195 1196 1197 1198 1199 1200 1201 1202 1203 1204 1205 1206 1207 1208 1209 1210 1211 1212 1213 1214 1215 1216 1217 1218 1219 1220 1221 1222 1223 1224 1225 1226 1227 1228 1229 1230 1231 1232 1233 1234 1235 1236 1237 1238 1239 1240 1241 1242 1243 1244 1245 1246 1247 1248 1249 1250 1251 1252 1253 1254 1255 1256 1257 1258 1259 1260 1261 1262 1263 1264 1265 1266 1267 1268 1269 1270 1271 1272 1273 1274 1275 1276 1277 1278 1279 1280 1281 1282 1283 1284 1285 1286 1287 1288 1289 1290 1291 1292 1293 1294 1295 1296 1297 1298 1299 1300 1301 1302 1303 1304 1305 1306 1307 1308 1309 1310 1311 1312 1313 1314 1315 1316 1317 1318 1319 1320 1321 1322 1323 1324 1325 1326 1327 1328 1329 1330 1331 1332 1333 1334 1335 1336 1337 1338 1339 1340 1341 1342 1343 1344 1345 1346 1347 1348 1349 1350 1351 1352 1353 1354 1355 1356 1357 1358 1359 1360 1361 1362 1363 1364 1365 1366 1367 1368 1369 1370 1371 1372 1373 1374 1375 1376 1377 1378 1379 1380 1381 1382 1383 1384 1385 1386 1387 1388 1389 1390 1391 1392 1393 1394 1395 1396 1397 1398 1399 1400 1401 1402 1403 1404 1405 1406 1407 1408 1409 1410 1411 1412 1413 1414 1415 1416 1417 1418 1419 1420 1421 1422 1423 1424 1425 1426 1427 1428 1429 1430 1431 1432 1433 1434 1435 1436 1437 1438 1439 1440 1441 1442 1443 1444 1445 1446 1447 1448 1449 1450 1451 1452 1453 1454 1455 1456 1457 1458 1459 1460 1461 1462 1463 1464 1465 1466 1467 1468 1469 1470 1471 1472 1473 1474 1475 1476 1477 1478 1479 1480 1481 1482 1483 1484 1485 1486 1487 1488 1489 1490 1491 1492 1493 1494 1495 1496 1497 1498 1499 1500 1501 1502 1503 1504 1505 1506 1507 1508 1509 1510 1511 1512 1513 1514 1515 1516 1517 1518 1519 1520 1521 1522 1523 1524 1525 1526 1527 1528 1529 1530 1531 1532 1533 1534 1535 1536 1537 1538 1539 1540 1541 1542 1543 1544 1545 1546 1547 1548 1549 1550 1551 1552 1553 1554 1555 1556 1557 1558 1559 1560 1561 1562 1563 1564 1565 1566 1567 1568 1569 1570 1571 1572 1573 1574 1575 1576 1577 1578 1579 1580 1581 1582 1583 1584 1585 1586 1587 1588 1589 1590 1591 1592 1593 1594 1595 1596 1597 1598 1599 1600 1601 1602 1603 1604 1605 1606 1607 1608 1609 1610 1611 1612 1613 1614 1615 1616 1617 1618 1619 1620 1621 1622 1623 1624 1625 1626 1627 1628 1629 1630 1631 1632 1633 1634 1635 1636 1637 1638 1639 1640 1641 1642 1643 1644 1645 1646 1647 1648 1649 1650 1651 1652 1653 1654 1655 1656 1657 1658 1659 1660 1661 1662 1663 1664 1665 1666 1667 1668 1669 1670 1671 1672 1673 1674 1675 1676 1677 1678 1679 1680 1681 1682 1683 1684 1685 1686 1687 1688 1689 1690 1691 1692 1693 1694 1695 1696 1697 1698 1699 1700 1701 1702 1703 1704 1705 1706 1707 1708 1709 1710 1711 1712 1713 1714 1715 1716 1717 1718 1719 1720 1721 1722 1723 1724 1725 1726 1727 1728 1729 1730 1731 1732 1733 1734 1735 1736 1737 1738 1739 1740 1741 1742 1743 1744 1745 1746 1747 1748 1749 1750 1751 1752 1753 1754 1755 1756 1757 1758 1759 1760 1761 1762 1763 1764 1765 1766 1767 1768 1769 1770 1771 1772 1773 1774 1775 1776 1777 1778 1779 1780 1781 1782 1783 1784 1785 1786 1787 1788 1789 1790 1791 1792 1793 1794 1795 1796 1797 1798 1799 1800 1801 1802 1803 1804 1805 1806 1807 1808 1809 1810 1811 1812 1813 1814 1815 1816 1817 1818 1819 1820 1821 1822 1823 1824 1825 1826 1827 1828 1829 1830 1831 1832 1833 1834 1835 1836 1837 1838 1839 1840 1841 1842 1843 1844 1845 1846 1847 1848 1849 1850 1851 1852 1853 1854 1855 1856 1857 1858 1859 1860 1861 1862 1863 1864 1865 1866 1867 1868 1869 1870 1871 1872 1873 1874 1875 1876 1877 1878 1879 1880 1881 1882 1883 1884 1885 1886 1887 1888 1889 1890 1891 1892 1893 1894 1895 1896 1897 1898 1899 1900 1901 1902 1903 1904 1905 1906 1907 1908 1909 1910 1911 1912 1913 1914 1915 1916 1917 1918 1919 1920 1921 1922 1923 1924 1925 1926 1927 1928 1929 1930 1931 1932 1933 1934 1935 1936 1937 1938 1939 1940 1941 1942 1943 1944 1945 1946 1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008

+

# THE NEW YORK PUBLIC LIBRARY

1250	1270	1290
------	------	------

DEFGYGHCSSESTNDKNEV

1310 1330 1350

MATCH WITH FIG. 2G

## FIG. 2G

MATCH WITH FIG. 2F

. . . . .  
 ATGCATTTCAGGACATTTCAATGAGTAATGTATCATGGAGAACTCTCAGACGGAATATA  
 -----+-----+-----+-----+-----+-----+-----+  
 TACGTAAAGTCCTGTAAAGTTACTCATTAACATAGTACCCCTCTTGAGAGCTGCGCTTATAT  
 A F Q D I S M S N V S W E N S Q T E Y S  
 1370 1390 1410

. . . . .  
 GTAAAACTTGTTTATAAGTTCGGTTAAGCACACCCAGTCAGAAAATGGCAATAAGACC  
 -----+-----+-----+-----+-----+-----+-----+  
 CATTTTGAACAAAATATCAAGGCAATTCGTGTGGTCAGTCTTTTACCGTTATTCTGG  
 K T C F I S S V K H T Q S E N G N K D H  
 1430 1450 1470

. . . . .  
 ATATAGATGAGAGTGGGAAATGAGGAAGACAGGCTCTTGAAAACCTCTTCGGAAATTT  
 -----+-----+-----+-----+-----+-----+-----+  
 TATATCTACTCTACCCCTTTTACTCCTTCTTCGTCCAGAACTTTTGAGAAGCCTTTAAA  
 I D E S G E N E E E A G L E N S S E I S  
 1490 1510 1530

. . . . .  
 CTGCAGATGAGTGGAGCAGGGGAAATATACTTAAAATTCAGTGGAGAGAAATATGAAC  
 -----+-----+-----+-----+-----+-----+-----+  
 GACGTACTACCTCGTCCCTTTTATATGAATTTTAAAGTCACCCCTCTCTTATAACTTG  
 A D E W S R G N I L K N S V G E N I E P  
 MATCH WITH FIG. 2H

17/41

SUBSTITUTE SHEET (RULE 26)

**MATCH WITH FIG. 2G**

1550 1570 1590  
CTGTGAAAATTTTAGTGCCCTGAAAAAAGTTTACCATGTAAAGTAAGTAATAATTAATATC  
GACACTTTTAAATCACGGACTTTTTCAAATGGTACATTTTCATTTCATTATTAATATAG  
V K I L V P E K S L P C K V S N N N Y P  
1610 1630 1650  
CAATCCCTGAACAAATGAATCTTAATGAAGATTTCATGTAAACAAAATCAAAATGTAATAG  
GTTAGGGACTTGTTACTTAGAATTACTTCTAAGTACATGTGTTTTTTTAGTTTACATTATC  
I P E Q M N L N E D S C N K K S N V I D  
1670 1690 1710  
ATAATAAATCTGGAAAAGTTACAGCTTATGATTTACTTAGCAATCGAGTAATCAAGAAAC  
TATTATTTAGACCTTTTCAATGTCGAATACTAAATGAATCGTTAGCTCATTAGTTCCTTG  
N K S G K V T A Y D L L S N R V I K K P  
1730 1750 1770  
CCATGTCAGCAAGTGCTCTTTTGTGTTCAAGATCATCGTCTCAGTTTCTCATAGAAAATC  
GGTACAGTCGTTACGAGAAAAACAAGTTCTAGTAGCAGGAGTCAAGAGTATCTTTTAG  
MATCH WITH FIG. 21

MATCH WITH FIG. 2H

FIG. 2I

M S A S A L F V Q D H R P Q F L I E N P  
 1790 1810 1830

CTAAGACTAGTTTAGAGGATGCAACACTACAAATTGAAGAACTGTGGAAGACATTGAGTG  
 -----+-----+-----+-----+-----+-----+-----+  
 GATTCTGATCAAAATCTCCTACGTTGTGATGTTTAACCTTCTTGACACCTTCTGTAACTCAC  
 K T S L E D A T L Q I E E L W K T L S E  
 1850 1870 1890

AAGAGGAAAACCTGAAATATGAAGAGAAGGCTACTAAAGACTTGGAAACGATACAAATAGTC  
 -----+-----+-----+-----+-----+-----+-----+  
 TTCTCCTTTTGACTTTATACTTCTCTCCGATGATTCTGAACCTTGCTATGTTATCAG  
 E E K L K Y E E K A T K D L E R Y N S Q  
 1910 1930 1950

AAATGAAGAGAGCCATTGAACAGGAGTCACAAATGTCACCTAAAGATGGCAGAAAAGA  
 -----+-----+-----+-----+-----+-----+-----+  
 TTACTTCTCTCGGTAACCTTGTCCTCAGTGTTTACAGTGATTTTCTACCGTCTTTTCT  
 M K R A I E Q E S Q M S L K D G R K I  
 1970 1990 2010

TAAACCCAGCGCATGGAATTGGCCCGAGAGCACAAGTTAAACCTCATATCTA  
 -----+-----+-----+-----+-----+-----+-----+  
 MATCH WITH FIG. 2J

19/41

SUBSTITUTE SHEET (RULE 26)

## FIG. 2J

MATCH WITH FIG. 2I

ATTTGGGTGGTCGGTACCTTAAACCGGCTTCGTGTTCAATTTTGAGTAATAGAT  
 K P T S A W N L A Q K H K L K T S L S N  
 2030 2050 2070

ATCAACCAAACTTGATGAACCTCCTTCAGTCCCAAATTGAAAAAAGAGAGTCAAAATA  
 -----+-----+-----+-----+-----+-----+-----+  
 TAGTTGGTTTTGAACTACTTGAGGAAGTCAGGTTTAACTTTTCTCCTCAGTTTAT  
 Q P K L D E L L Q S Q I E K R R S Q N I  
 2090 2110 2130

TTAAAATGGTACAGATCCCTTTTCTATGAAAACTTAAATAAATTTAAGAAACAAA  
 -----+-----+-----+-----+-----+-----+-----+  
 AATTTACCATGTCTAGGGGAAAGATACTTTTGAATTTTATTTAAATCTTTGTTT  
 K M V Q I P F S M K N L K I N F K K Q N  
 2150 2170 2190

ACAAAGTTGACTTAGAAGAGAGGATGAACCTTGCTTGATCCACAATCTCAGGTTTCCTG  
 -----+-----+-----+-----+-----+-----+-----+  
 TGTTTCAACTGAATCTCTCTCCTACTTGGAAACGAACTAGGTGTAGAGTCCAAGGAC  
 K V D L E E K D E P C L I H N L R F P D  
 2210 2230 2250

MATCH WITH FIG. 2K

20/41

SUBSTITUTE SHEET (RULE 26)

## FIG. 2K

MATCH WITH FIG. 2J

ATGCATGGCTAATGACATCCAAACAGAGGTAATGTTATTAAATCCATATAGAGTAGAAG  
 -----+-----+-----+-----+-----+-----+-----+  
 TACGTACCGATTACTGTAGGTTTGTCTCCATTACAATAATTAGGTATATCTCATCTTC  
 A W L M T S K T E V M L L N P Y R V E E  
 2270 2290 2310

AAGCCCTGCTATTAAAGACTTCTTGAGAAATCATAAACTTCCTGCAGAGCCACTGGA  
 -----+-----+-----+-----+-----+-----+-----+  
 TTCGGGACGATAAATTTCTGAAGAACTCTTAGTATTGAAGGACGTCTCGGTGACCTTT  
 A L L F K R L L E N H K L P A E P L E K  
 2330 2350 2370

AGCCAATTATGTTAACAGAGAGTCTTTTTTAATGGATCTCATTTATTAGACGTTTATATA  
 -----+-----+-----+-----+-----+-----+-----+  
 TCGGTTAATACAATTGTCTCTCAGAAAAATTACCTAGAGTAATAAATCTGCAGAAATATAT  
 P I M L T E S L F N G S H Y L D V L Y K  
 2390 2410 2430

AAATGACAGAGATGACCAAGATACAGTGGATCAACTTACCTGTCTGATCCTCGTCTTA  
 -----+-----+-----+-----+-----+-----+-----+  
 TTTACTGTCGTCTACTGGTTTCTATGTACCTAGTTGAATGACAGACTAGGAGCAGAAT  
 M T A D D Q R Y S G S T Y L S D P R L T  
 2450 2470 2490

MATCH WITH FIG. 3L

21/41

SUBSTITUTE SHEET (RULE 26)

MATCH WITH FIG. 2K      FIG. 2L

```

CAGCGAATGGTTTCAAGATAAAATTGATACCAGGAGTTTCAATTACTGAAATTAATTGG
-----+-----+-----+-----+-----+-----+-----+
GTCGCTTACCAAAGTTCTATTTAACTATGGTCCTCAAGTTAATGACTTTTAATGAACC
A N G F K I K L I P G V S I T E N Y L E
2510                               2530                2550

AAATAGAGGAATGGCTAATTGTCTCCCATTTCTATGGAGTAGCAGATTAAAGAAATTC
-----+-----+-----+-----+-----+-----+-----+
TTTATCTTCCTTACCGATTAAACAGAGGGTAAGATACCTCATCGTCTAAATTTCTTTAAG
I E G M A N C L P F Y G V A D L K E I L
2570                               2590                2610

TTAATGCTATATTAAACAGAAATGCAAAGGAAGTTTATGAATGTAGACCTCGCAAAGTGA
-----+-----+-----+-----+-----+-----+-----+
AATTACGATATAATTGTCTTTACGTTTCCCTTCAATACTTACATCTCGAGCGTTTCACT
N A I L N R N A K E V Y E C R P R K V I
2630                               2650                2670

TAAGTTATTTAGAGGAGAACGAGTGGCTCTATCCAGACAATTACCCATGTACTATCAA
-----+-----+-----+-----+-----+-----+-----+
ATTCAATAAATCTCCCTCTTCGTACGCAGATAGGTCTGTAAATGGTACATGAATAGTT
S Y L E G E A V R L S R Q L P M Y L S K
2690                               2710                2730

```

MATCH WITH FIG. 2M

22/41

SUBSTITUTE SHEET (RULE 26)



MATCH WITH FIG. 2L

```

AAGAGGACATCCAAGACATTATCTACAGAATGAAGCACCAGTTTGGAAATGAAATTAAAG
-----+-----+-----+-----+-----+-----+-----+
TTCTCCTGTAGGTTCTGTAATAGATGTCTTACTTCGTGTCACAAACCTTTACTTTAATTTC
E D I Q D I I Y R M K H Q F G N E I K E
2750 2770 2790

AGTGTGTTCA TGGTCGCCCATTTTTCATCATTTTAACCTATCTTCCAGAACTACATGAT
-----+-----+-----+-----+-----+-----+-----+
TCACACAAGTACCAGCGGTAAAGTAGTAATAATTGGATAGAGGTCTTTGATGTACTA
C V H G R P F F H H L T Y L P E T
2810 2830 2850

TAAATATGTTTAAAGAAGATTAGTTACCATTGAAATTTGGTTCTGTCTATAAACAGCATGAG
-----+-----+-----+-----+-----+-----+-----+
ATTATACAAATTCTTCTAATCAATGGTAACTTTAACCAAGACAGATATTTTGTCTGCTACTC
2870 2890 2910

TCTGGTTTAAATTATCTTTGTATTATGTGTCACATGGTTATTTTAAATGAGGATTCA
-----+-----+-----+-----+-----+-----+-----+
AGACCAAAATTTAATAGAAACATAATACACAGTGTACCAATAAAATTTACTCCTAAGT
2930 2950 2970

CTGACTTGTTTTATATTGAAAAAGTTCCACGTATTGTAGAAACGTAAATAAACTAAT
-----+-----+-----+-----+-----+-----+-----+
GACTGAACAAAATATAACTTTTTCAGGTGCATAACATCTTTTGCATTATTGATTTA
AAC
TTG

```

FIG. 2M

FIG. 3A

CGAGGCGGATCGGGTGTTCATCCATGGAGCGAGCTGAGAGCTCGAGTACAGAACCTGCT  
-----+-----+-----+-----+-----+-----+  
GCTCCGCCCTAGCCCCACAACGTAGGTACCTCGCTCGACTCTCGAGCTCATGTCTTGACGA

M E R A E S S T E P A

40                  60                  80

. . . . .  
 AAGGCCATCAACCTATTGATCGGAAGTCAGTCCATCAGATTTCCTCTGCGGCAGGTGGTA  
 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----  
 TTCCGGTAGTTTGGATAACTAGCCTTCAGTCAGGTCAGTCTAAACGAGACCCGTCCACCAT  
 K A I K P I D R K S V H Q I C S G Q V V  
 100 120 140

CTGAGTCTAAGCACTGCCGTAAAGGAGTTAGTAGAAAACAGTCTGGATGCTGGTGCCACT  
-----+-----+-----+-----+-----+-----+-----  
GACTCAGATTCTGTACGCCAATTTCCTCAATCATCTTTTGTACAGACCTACGACCACGGTGA  
L S L S T A V K E L V E N S L D A G A T  
160 180 200

AATATTGATCTAAAGCTTAAGGACTATGGAGTGGATCTTATTGAAGTTTCAGACAAATGGA  
-----+-----+-----+-----+-----+-----+  
TTATACTAGATTTTCGAATTCCCTGATACCTCACCTAGAATAAAGTCTGTTCCT

MATCH WITH FIG. 3B

**MATCH WITH FIG. 3B**

MATCH WITH FIG. 3A      **FIG. 3B**

```

N I D L K L K D Y G V D L I E V S D N G
220 . 240 260
TGTGGGTAAGAAGAAACTTCGAAGGCTTAACTCTGAAACATCACACATCTAAGATT
-----+-----+-----+-----+-----+-----+-----+
ACACCCATCTTCTTCTTTGAAGCTTCGGAATTGAGACTTTGTAGTGTGTAGATTCTAA
C G V E E E N F E G L T L K H H T S K I
280 300 320
CAAGAGTTGCCGACCTAACTCAGGTGAAACTTTTGGCTTTTCGGGGGAAGCTCTGAGC
25 340 360 380
GTCTCAAACGGCTGGATTGAGTCCAACTTTGAAACCGAAAGCCCCCTTCGAGACTCG
Q E F A D L T T Q V E T F G F R G E A L S
41 340 360 380
TCACTTTGTGCACTGAGCGATGTCAACCATTTCTACCTGCCACGCAATCGCGAAGGTTGGA
-----+-----+-----+-----+-----+-----+-----+
AGTGAAACACGTCGCTACAGTGGTAAAGATGGACGGTGGTAGCCGCTTCCCAACCT
S L C A L S D V T I S T C H A S A K V G
400 420 440
ACTCGACTGATGTTTGATCACAAATGGGAAATTAATCCAGAAACCCCTACCCCCGCCCC
-----+-----+-----+-----+-----+-----+-----+

```

MATCH WITH FIG. 3C

MATCH WITH FIG. 3B FIG. 3C

```

TGAGCTGACTACAACTAGTGTACCCCTTTAATAGTCTTTTGGGGGATGGGGCGGGGG
T R L M F D H N G K I I Q K T P Y P R P
460                               480                               500

AGAGGGACCACAGTCAGCGTGCAGCAGTTATTTCCACACTACCTGTGCGCCATAAGGAA
-----+-----+-----+-----+-----+-----+-----+-----+
TCTCCCTGGTGCAGTCGCACGTCGTCAATAAAAGGTGTGATGGACACGCGGTATTCCTT
R G T T V S V Q Q L F S T L P V R H K E
520                               540                               560

TTTCAAAGGAATATTAAGAAGGAGTATGCCAAAATGGTCCAGGTCTTACATGCATACTGT
-----+-----+-----+-----+-----+-----+-----+-----+
AAAGTTTCCTTATAATCTTCCCTCATACGGTTTACCAGGTCCAGAAATGTACGTATGACA
F Q R N I K K E Y A K M V Q V L H A Y C
580                               600                               620

ATCATTTTCAGCAGGCATCCGTGTAAAGTTGCACCAATCAGCTTGGACAAGGAAACGACAG
-----+-----+-----+-----+-----+-----+-----+-----+
TAGTAAAGTCGTCGTCAGCACATTCAACGTGTTAGTCGAACCTGTTCCCTTTTGCTGTC
I I S A G I R V S C T N Q L G Q G K R Q
640                               660                               680

CCTGTGGTATGCACAGGTGGAAGCCCCAGCATAAAGGAAAATATCGGCTCTGTGTTGGG
MATCH WITH FIG. 3D

```

## FIG. 3D

MATCH WITH FIG. 3C

```

+-----+-----+-----+-----+-----+-----+-----+-----+
GGACACCATACGTGTCCACCTTCGGGGTCGTATTTCCTTTTATAGCCGAGACACAAACCC
P V V C T G G S P S I K E N I G S V F G
700 720 740

CAGAAGCAGTTGCAAGCCTCATTCCTTTTGTTCAGCTGCCCCCTAGTCACTCCGTGTGT
+-----+-----+-----+-----+-----+-----+-----+-----+
GTCTTCGTCAACGTTTCGGAGTAAGGAAACAAGTCGACGGGGATCACTGAGGCACACA
Q K Q L Q S L I P F V Q L P P S D S V C
760 780 800

GAAGAGTACGGTTTGAGCTGTTTCGGATGCTCTGCATAATCTTTTACATCTCAGGTTTC
+-----+-----+-----+-----+-----+-----+-----+-----+
CTTCTCATGCCAAACTCGACAAGCCTACGAGACGTATTAGAAAATGTAGAGTCCAAAG
E E Y G L S C S D A L H N L F Y I S G F
820 840 860

ATTTCACAAATGCACGCATGGAGTTGGAGGAGTTCAACAGACAGACAGTTTTCTTTATC
+-----+-----+-----+-----+-----+-----+-----+-----+
TAAAGTGTTACGTGCGTACCTCAACCTTCCTCAAGTTGTCTGTCTGTCAAAAAGAAATAG
I S Q C T H G V G R S S T D R Q F F I
880 900 920

```

MATCH WITH FIG. 3E

## FIG. 3E

MATCH WITH FIG. 3D

AACCGCGGCCCTTGTGACCCAGCAAGGTCTGCAGACTCGTGAAATGAGGTCTACCCACATG  
 TTGGCCGCCGGAACACTGGGTGCTTCCAGACGCTCTGAGCACTTACTCCAGATGGTGATC  
 N R R P C D P A K V C R L V N E V Y H M  
 940 960 980

TATAATCGACACCAGTATCCATTGTGTTCTTAACATTTCTGTGATTTCAGAAATGCGTT  
 ATATTAGCTGTGTCATAGGTAACAACAAGAATTGTAAGACAACTAAGTCTTACGCAA  
 Y N R H Q Y P F V V L N I S V D S E C V  
 1000 1020 1040

GATATCAATGTTACTCCAGATAAAGGCAAAATTTGCTACAAGAGGAAAGCTTTTGTG  
 CTATAGTTACAATGAGGTCTATTTCCGTTTAAACGATGTTCTCCTTTTCGAAACAAC  
 D I N V T P D K R Q I L L Q E E K L L L  
 1060 1080 1100

GCAGTTTAAAGACCTCTTTGATAGGAATGTTGATAGTGATGTCACAAGCTAAATGTC  
 CGTCAAAATTTCTGGAGAAACTATCCTTACAACACTATCACTACAGTTGTTTCGATTACAG  
 A V L K T S L I G M F D S D V N K L N V  
 MATCH WITH FIG. 3F

29741

**MATCH WITH FIG. 3F**

K P H S P K T P E P R R S P L G Q K R G  
1360 1380 1400  
ATGCTGTCTTCTAGCACTTCAGGTGCCATCTCTGACAAAGCGCTCCTGAGACCTCAGAAA  
TACGACAGAAGATCGTGAAGTCCACGGTAGAGACTGTTCGCGAGGACTCTGGAGTCTTT  
M L S S S T S G A I S D K G V L R P Q K  
1420 1440 1460  
GAGGCAGTGAGTTCAGTCAAGGACCCAGTGACCCTACGGACAGAGCGGAGGTGGAGAAG  
CTCCGTCACCTAAGGTCAAGTCCCTGGGTCACTGGGATGCCGTGTCTGCCCTCCACCTCTC  
E A V S S S H G P S D P T D R A E V E K  
1480 1500 1520  
GACTCGGGGCACGGCAGCACTTCCTCGTGATCTCTGAGGGTTTCAGCATCCAGACACGGGC  
CTGAGCCCCGTGCCGTGTAAGGCACCTAAGACTCCCCAAGTCGTAGGGTCTGTGCCCG  
D S G H G S T S V D S E G F S I P D T G  
1540 1560 1580  
AGTCACTGCAGCAGGAGTATGCGGCCAGCTCCCCAGGGGACAGGGGCTCGCAGGAACAT

**MATCH WITH FIG. 3H**



**MATCH WITH FIG. 3G**

GTGGACTCTCAGGAGAAAGCGCCCTGAACCTGACGACTCTTTTCAGATGTGGACTGCCCAT  
-----+-----+-----+-----+-----+-----+  
CACCTGAGAGTCCCTCTTTCGGCGACTTGTGCTGAGAAAAGTCTACACCTGACGGTA  
V D S Q E K A P E T D D S F S D V D C H  
1660 1680 1700

TCAAACCAGGAATACCGGATGTAATTTTCGAGTTTTCCTCAGCCAACTAATCTCGCA  
 AGTTTGGTCCTTCTATGGCCTACATTAAAGCTCAAACGGAGTCGGTTGATTAGAGCGT  
 S N Q E D T G C K F R V L P Q P T N L A  
 1720 1740 1760

ACCCCAAACACAAAGCGTTTAAATAAAGAGAAATCTTTCCAGTTCTGACATTTGTCAA  
 TGGGGTTTGTGTTCGCCAAAATTTTTCCTCTTTAAGAAAGGTCAAGACTGTAAACAGTT  
 T P N T K R F K K E E I L S S S D I C Q  
 1780 1800 1820

MATCH WITH FIG. 3I

MATCH WITH FIG. 3H      **FIG. 3I**

```

AAGTTAGTAAATACTCAGGACATGTTCAGCCTCTCAGGTTGATGTAGCTGTGAAATTAAT
-----+-----+-----+-----+-----+-----+-----+-----+
TTC AATCATTTATGAGTCCGTGACAGTCGGAGAGTCCAACTACATCGACACTTTTAAATTA
K L V N T Q D M S A S Q V D V A V K I N
1840                               1860                               1880

AAGAAAGTTGTGCCCCCTGGACTTTTCTATGAGTCTTTAGCTAAACGAATAAAGCAGTTA
-----+-----+-----+-----+-----+-----+-----+-----+
TTC TTTCAACACGGGACCTGAAAAGATACTCAAGAAATCGATTGCTTATTTTCGTCAAT
K K V V P L D F S M S S L A K R I K Q L
1900                               1920                               1940

CATCATGAAGCACAGCAAGTGAAGGGGAACAGAAATTACAGGAAGTTTAGGGCAAAGATT
-----+-----+-----+-----+-----+-----+-----+-----+
GTAGTACTTCGTGTCGTTTCACTTCCCTTGCTCTTAATGTCCTTCAAATCCCCTTCTAA
H H E A Q Q S E G E Q N Y R K F R A K I
1960                               1980                               2000

TGTCCTGGAGAAATCAAGCAGCCGAAGATGAAC TGAAGAAAGAGATAAGTAAACGATG
-----+-----+-----+-----+-----+-----+-----+-----+
ACAGGACCTCTTTAGTTCGTCGGCTTCTACTTGATTCCTTTTCTCTATTCATTTTGCTAC
C P G E N Q A A E D E L R K E I S K T M
2020                               2040                               2060

```

MATCH WITH FIG. 3J

32/41

SUBSTITUTE SHEET (RULE 26)

## FIG. 3J

MATCH WITH FIG. 3I

TTTGCAGAAATCGAAATCATTTGGTCAGTTTAACCTGGGATTATATAACCAAACTGAAT  
 AAACGCTCTTACCTTTAGTAACCAAGTCAGTCAAATTGGACCCCTAAATATTATGTTGACTTA  
 F A E M E I I G Q F N L G F I I T K L N

2080 2100 2120

GAGGATATCTTCATAGTGGACCAAGCATGCCACGGACGAGAAGTATAACTTCGAGATGCTG  
 CTCCTATAGAAGTATCACCTGGTCGTACGGTGCCTGCTCTTCATATTGAAGCTCTACGAC  
 E D I F I V D Q H A T D E K Y N F E M L  
 2140 2160 2180

CAGCAGCACACCGTGCTCCAGGGCAGAGGCTCATAGCACCTCAGACTCTCAACTTAACT  
 GTCGTCGTGGCAGAGGTCCCGTCTCCGAGTATCGTGGAGCTCGAGAGTTGAATTGA  
 Q Q H T V L Q G Q R L I A P Q T L N L T  
 2200 2220 2240

GCTGTTAATGAAGCTGTTCTGATAGAAAATCTGGAAATATTTAGAAAGAATGGCTTTGAT  
 CGACAATTACTTCGACAAGACTATCTTTTAGACCTTTTATAAATCTTTCTTACCGAAACTA  
 A V N E A V L I E N L E I F R K N G F D

MATCH WITH FIG. 3K

33/41

SUBSTITUTE SHEET (RULE 26)

## FIG. 3K

MATCH WITH FIG. 3J

```

2260      .      .      .      .      .      .      .      .      .      .
      TTTGTTATCGATGAAAATGCTCCAGTCACTGAAAGGGCTAAACTGATTTCCCTTGCCAACT
      -----+-----+-----+-----+-----+-----+-----+-----
      AACAAATAGCTACTTTTACGAGGTCAGTCACTTTCCCGATTGACTAAAGGAACGGTTGA
      F V I D E N A P V T E R A K L I S L P T
      2320      2340      2360

      .      .      .      .      .      .      .      .      .      .
      AGTAAAACTGGACCTTCGGACCCAGGACGTCGATGAAGTCTTTCATGCTGAGCGAC
      -----+-----+-----+-----+-----+-----+-----+-----
      TCAATTTTGACCTGGAAGCCTGGGTCCTGCAGCTACTTACTAGAAAGTACGACTCGCTG
      S K N W T F G P Q D V D E L I F M L S D
      2380      2400      2420

      .      .      .      .      .      .      .      .      .      .
      AGCCCTGGGGTCATGTGCCGGCCTTCCCGAGTCAAGCAGATGTTTGCCCTCCAGAGCCCTGC
      -----+-----+-----+-----+-----+-----+-----+-----
      TCGGACCCAGTACACGGCCGGAAGGGCTCAGTTCGTCTACAAACGGAGGTCTCGGACG
      S P G V M C R P S R V K Q M F A S R A C
      2440      2460      2480

      .      .      .      .      .      .      .      .      .      .
      CGGAAGTCGGTGATGATTGGGACTGCTCTTAACACAAGCGAGATGAAGAACTGATCACC
      -----+-----+-----+-----+-----+-----+-----+-----
      GCCTTCAGCCACTACTAACCCCTGACGAGAAATGTGTTGCTCTACTTCTTTGACTAGTGG
      MATCH WITH FIG. 3L

```

34/41

SUBSTITUTE SHEET (RULE 26)

MATCH WITH FIG. 3K

R	K	S	V	M	I	G	T	A	L	N	T	S	E	M	K	K	L	I	T
2500						2520											2540		
CACATGGGGAGATGGACCA	CCCTGGAACTGTCCCATGGAAGGCCAACCATGAGACAC																		
-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+
GTGTACCCCTCTACCTGGTGGGACCTTGACAGGGGTACCTTCCGGTTGGTACTCTGTG																			
H	M	G	E	M	D	H	P	W	N	C	P	H	G	R	P	T	M	R	H
2560						2580													
ATCGCCAACCTGGGTGTCATTTCTCAGAACTGACCGTAGTCACTGTATGGAATAATTGGT																			
-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+
TAGCGGTTGGACCCACAGTAAAGAGTCTTGACTGGCATCAGTGACATACCTTATTAAACCA																			
I	A	N	L	G	V	I	S	Q	N	*									
2620						2640													
TTTATCGCAGATTTTATGTTTTGAAAGACAGAGTCTTCACTAAACCTTTTGTGTTTAA																			
-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+
AAATAGCGTCTAAAAAATACAAAACCTTTCTGTCTCAGAAAGTGATTTGGAAAAACAAATTT																			
2680						2700													
ATGAAACCTGCTACTTAAAAAATAACATCACATCACACCCATTAAAGTGATCTTGAGAAC																			
-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+	-----+
TACTTTGGACGATGAATTTTATGTGTAGTGTGGGTAAATTTTCACTAGAACTCTTG																			
2740																			
CTTTTCAAACC																			
-----+																			
GAAAGTTTGG																			

FIG. 3L

## FIG. 4A

yPMS1	tfhhienllietekrckqkeqryipvkylfsmtqIH
hMLH2	-----MK
hMLH3	teraessstepaka-----IK
yPMS1	YGLESTIECSDNCGDIDPSNYEFLALKHYTSKIAKFQ
hMLH2	YGFDKIEVRDNGEGIKAVDAPVMAMKYYTSKINSHE
hMLH3	YGVLLIEVSDNCGGVVEENFEGTLKHHTSKIQEFA
yPMS1	GHITSKTTTSRNKGTTVLVSQLEHNL PVRQKEFSKT
hMLH2	GHILSOKPSHLGQGTTVTALRLEKNLPVRKQFYSTA
hMLH3	GKIIQKTPYPRPRGTTVSVQOLESTLPVRHKEFORN
yPMS1	ssmrknissvfgaggmrgleevdlvldlnpfknrm
hMLH2	kmalmvlgtavmnnmesfqyhseesqiylsgflpk
hMLH3	psikenigsvfgqkqlqslipfvqlppsdsvceeyg
yPMS1	PVEYSTLLKCCNEVYKTFnnvq----FPAVFLNLEF
hMLH2	PVHQKDILKLIRHHYNLkclkestrlyPVFFLKIDV
hMLH3	PCDPAKVCLVNEVHMynrho----YPFVVLNLSV
yPMS1	krmcsqseqqaqkrllktevfdörstthesdnenyht
hMLH2	yennktdvsaadivlsktaetdvlnkvessgknys
hMLH3	vsqqplldvegnlikmhaadlekpmvekqdqpslr
yPMS1	secevsdssvvldegnsstptkklpsiktdsqnl
hMLH2	snidkntknafqdismsnvswensqteysktcfiss
hMLH3	gmlssstsgaisdkgvlrpqkeavssshgpsdptdr
yPMS1	avlsqadglvfvadnechehtndcchqerrgstdeq
hMLH2	nsvgeniepvkilvpekslpckvsnnnypipeqmnl
hMLH3	hvdsgkapetddsfsdvchsnqedtgckfrvlpq

MATCH WITH FIG. 4B

MATCH WITH FIG. 4C

36/41

SUBSTITUTE SHEET (RULE 26)

## FIG. 4B

MATCH WITH FIG. 4A	QINDIDVHRITSGQVITDLTTAVKELVDNSIDANANQTEIIFKD	80
	QLPAATVRLLSSSQIITSVVSVKELIENSLDAGATSVQVLEN	46
	PIDRKSVHQICSGQVLSLSTAVKELVENS LDAGATNIDLKLKD	60
	DVAKVOTLGFGEALSSLCGIAKLSVITTTSPPK-ADKLEYDMV	159
	DLENLTITYGFRGEALGSICCIAEVLITRTAADNFSTQYVLDGS	126
	DLTQVETFGFRGEALSSLCALSDVTHSTCHASAKVGTRLMFDHN	140
	fkqrftkcltviqgyaiaaiaikfsvwnitpkqkknllstmrn	239
	klckdeikkigdllmsfgilkpdlrivfvhnkaviwqksrvsdh	206
	ikkeyakmvqvlhayciaisagirvscnqlgqgkrqpvvctggs	220
	gkytdpdfldldykirvkgysisqnsfgcgrNSKDROFIYVNR	319
	cdadhsftsl-----STPERSFIFINR	265
	lscsdalhnlfyisgffisqcthgvr-----SSTDROFFFINR	295
	PMSLIDVNVTPDKRVILLHNERAVIDIFKTTLSDYYNrqelalp	395
	PTADVNVNLTDPDKSQVLLQNKESVLI ALENLMTTCYGplpstns	345
	DSECVDINVTDPDKRQILLQEEKLLIAVLKTSLIGMEDsdvnkln	371
	arsesnqsnhahfnsttgvidksngteltsvmdgnytnvtdvig	475
	nvdtsvipfqndmhndesgkntddclnhqisigdfgyghcssei	425
	tgeekkdvsisrlreafslrhttenkphspktpeprsrplgqkr	451
	dlnlnnfsnpefqnitspdkarslekvveepvyfdidgekfqek	555
	vkhtqsengnkdhidesgeneeeeaglenesseisadewsergnlk	505
	aevekdsghgstsvdsegfsipdtgshcsseyaasspgdrqsqe	531
	ddeadsiyaeiepveinvrtplknsrksiskdnysrlsdglthr	635
	neoscnnkksnvidnksqkvtaydllsnrvikkpmsasalvqdh	585
	ptnlatpntkrfkkeeilsssdicqklvntqdmssasqvdvavki	611

MATCH WITH FIG. 4D

37/41

SUBSTITUTE SHEET (RULE 26)

## FIG. 4C

MATCH WITH FIG. 4A

yPMS1 kfedeileynlstknfkeiskngkqmsiiskrkse  
 hMLH2 rpqflienkptsledatlqieelwktlseeeklkye  
 hMLH3 nkkvvplofsmsslakrikqlhheaqqsegeqnyrk

yPMS1 iivtrkvdnksdifiyd sdekynfetlqavtvf  
 hMLH2 hklktsisnqpkldellqsqiekrrsqnikmvqipf  
 hMLH3 nedifiydgahatdekynfemlgghtvlggqrliapq

yPMS1 srvkllslptskqtlfdlgdfnelihlikedgglrr  
 hMLH2 llnpyrveeallfkrllenhklpaepklekpimltes  
 hMLH3 tsknwtfgpqdvdelifmlsdspgvmc-----

yPMS1 -----  
 hMLH2 vsitenyleiegmancplptygvadlkeilnailnrn  
 hMLH3 -----

yPMS1 eldkpw--NCPHGRPTMRHLMELrdwssfskdyei  
 hMLH2 hqfgneikeCVHGRPFHFHLLTYLpett-----  
 hMLH3 emdhpw--NCPHGRPTMRHIANLgvisgn-----

MATCH WITH FIG. 4D

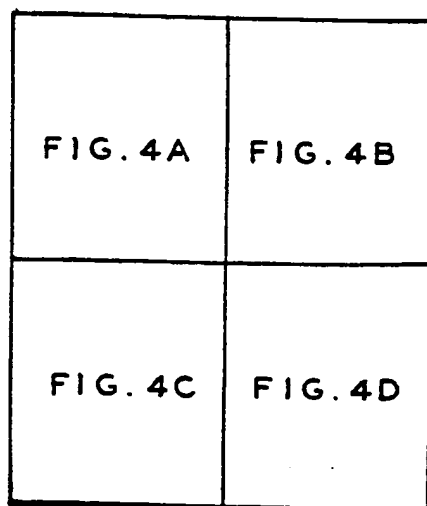


FIG. 4



## FIG. 4D

## MATCH WITH FIG. 4B

aqeniiknkdeledfeggekyltltvskndfkkmevvqgfnlgf 715  
 ekatkdlerynsqmkraiegesqmslkdgrkkikptsawnlaqk 665  
 frakicpgenqaaedelrkeisktmfaemeiigqfnlgfiitkl 691

MATCH WITH FIG. 4C

ksqkliipqpvelsvidelvvldnlpvfekngfklkideeeefg 795  
 smknlnkinfkkqnkvdleekdepclihnlrfdawlmtsktevm 745  
 tlnltavneavlienleifrknngfdvidenapvteraklislp 771

dni----- 834  
 lfngshyldvlykmtaddqrysgstylsdprltangfkiklipg 825  
 ----- 798

-----RCSKIRSMFAMRACRSSIMIGKPLNKKTMTRVHNLS 871  
 akevyecPPRKVISYLEGEAVRLSRQLPMYLSKEDQODIITYRMK 905  
 -----PPSRVKQMEASRACRKSVMIGTALNTSEMKKLITHMG 835

904  
 932  
 862

FIG. 6A

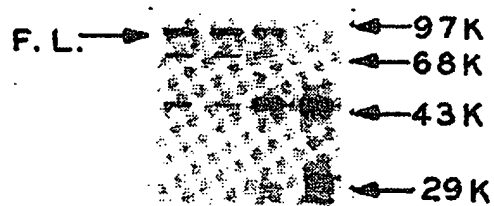
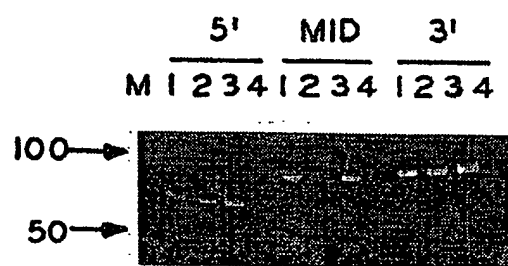


FIG. 6B



41/41

SUBSTITUTE SHEET (RULE 26)

## INTERNATIONAL SEARCH REPORT

International application No.  
PCT/US95/01035

**A. CLASSIFICATION OF SUBJECT MATTER**

IPC(6) : C12Q 1/68; C12N 9/08; A61K 51/00; C07K 1/00

US CL : Please See Extra Sheet.

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

U.S. : Please See Extra Sheet.

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Molecular and Cellular Biology, Volume 14, Number 1, issued January 1994, Prolla et al, "Dual requirement in Yeast DNA mismatch repair for MLH1 and PMS1, Two homologs of the Bacterial mutL gene," pages 407-415, especially page 407, column 2, line 3.	1-18, 21-23
P,Y	American Journal of Human Genetics, Volume 55, issued July 1994, Nystrom-Lahti et al, "Mismatch repair genes on Chromosome 2p and 3p account of a major share of Hereditary Nonpolyposis Colorectal Cancer families evaluable by linkage", pages 659-665, especially page 663, column 1, lines 9-13, and page 664, column 1, lines 19-30.	1-18, 21-23, 28-31

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

•	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance		
"E"	earlier document published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search

03 MAY 1995

Date of mailing of the international search report

22 MAY 1995

Name and mailing address of the ISA/US  
Commissioner of Patents and Trademarks  
Box PCT  
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

Dianne Rees, Ph.D.

Telephone No. (703) 308-0196

## INTERNATIONAL SEARCH REPORT

International application No.  
PCT/US95/01035

## C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P, Y	Science, Volume 265, issued August 1994, Prolla et al, "MLH1, PMS1 and MSH2 interactions during the initiation of DNA mismatch repair in yeast", pages 1091-1093, especially page 1091, column 1, and column 3, lines 1-5.	19, 20, 24-27
P, X ----- P, Y	Science, Volume 263, issued 18 March 1994, Papadopoulos et al, "Mutation of a <i>mutL</i> homolog in Hereditary Colon Cancer", pages 1625-1629, especially page 1626, column 1, paragraphs 1 and 2, figure 1, figure 3, and page 1627, column 3, paragraph 2, and p1628, notes: 13, 16, 17, 18, 20, 24, 25, 27.	5, 21-23, 29-31 ----- 1-4, 6-20, 24-28
P, X ----- P, Y	Nature, Volume 368, issued 17 March 1994, Bronner et al, "Mutation in the DNA mismatch repair gene homologue <i>hMLH1</i> is associated with hereditary non-polyposis colon cancer, pages 258-261, especially page 259, figure 1, page 260, figure 2 and 3.	21, 31 ----- 1-11, 14, 16-18, 25, 28-30
P, X ----- P, Y	Biochemical and Biophysical Research Communications, Volume 204, Number 3, issued 15 November 1994, Horii et al, "Cloning, Characterization and Chromosomal assignment of the human genes homologous to <i>PMS1</i> , a member of mismatch repair genes, pages 1257-1264, especially, page 1257, abstract, lines 10-14, page 1261, figure 2, and page 1262, figure 3.	7, 10, 13, 21, 31 ----- 1-6, 8, 9, 11, 12- 20, 22-30
Y	Cell, Volume 75, issued 16 December 1993, Leach et al, "Mutations of <i>mutS</i> homolog in Hereditary Nonpolyposis Colorectal Cancer", pages 1215-1225, especially page 1219, column 1, paragraph 3.	31

# INTERNATIONAL SEARCH REPORT

International application No.  
PCT/US95/01035

## A. CLASSIFICATION OF SUBJECT MATTER:

US CL :

435/6, 192.1, 193.1; 530/300, 350, 358, 387.3, 388.21; 536/23.1, 23.4, 24.31

## B. FIELDS SEARCHED

Minimum documentation searched

Classification System: U.S.

435/6, 192.1, 193.1; 530/300, 350, 358, 387.3, 388.21; 536/23.1, 23.4, 24.31

## B. FIELDS SEARCHED

Electronic data bases consulted (Name of data base and where practicable terms used):

BIOSIS, MEDLINE, EMBASE, CAPLUS, HCA, USPATFULL, WPIDS, CANCERLIT, GENBANK, GENBANK, GENBANK-NEW, UEMBL (searched on seq IDs from related US case, US08187757, CRF disk was defective))  
Search terms: human DNA repair (genes or proteins), mutator genes, mutL, hMLH1, hMLH2, hMLH3, colon cancer, microsatellite instability, Haseltine, Prolla, Liskay

## BOX II. OBSERVATIONS WHERE UNITY OF INVENTION WAS LACKING

This ISA found multiple inventions as follows:

- I. Claims 1-23, drawn to polynucleotides encoding polypeptides having the deduced amino acid sequences of hMLH-encoded proteins, their analogs or derivatives, vectors containing said polynucleotides, host cells genetically engineered with said vectors, process of growing said host cells.
- II. Claims 24-27, drawn to polypeptides and methods of polypeptide production from host cells expressing hMLH genes.
- III. Claims 28-31, drawn to a process for diagnosing cancer susceptibility comprising identifying mutations in hMLH1, hMLH2, hMLH3 and the human homolog of bacterial mutL.

An Election of Species for Groups I, II, and III is required wherein:

species A is drawn to hMLH1

species B is drawn to hMLH2

species C is drawn to hMLH3

and wherein Group III has an additional species:

species D, drawn to the human homolog of bacterial mutL.

These groups are separate and distinct from each other. Group I is drawn to products which are polynucleotides while Group II is drawn to products which are polypeptides and to a process of making said polypeptides. The products of Groups I and II have different structural and biochemical properties and may be used in distinctly different processes. Polynucleotides may be used as probes in linkage analyses, and DNA-based genetic therapy while polypeptides may be used in protein-based therapies. While the product Group I is linked to the process of Group II these do not share a common special technical feature according to PCT Rule 13.2 as "analogs, derivatives and variants" of group I are known in the art (Horii et al, Biochem. Biophys. Res. Commun., 28 November 1994). For the same reasons the product of Group I is also not technically linked to the process of Group III.

Species A-C (Groups I and II) and A-D (Group II) do not relate to a single inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2 "the commonly shared structure" does not "constitute a structurally distinctive portion in view of the prior art", i.e. in view of Horii et al. 1994. Further the nonobvious differences in sequence structures between these genes render these genes structurally and functionally distinct. Accordingly, the claims are not so linked by a special technical feature within the meaning of PCT Rule 13.2 so as to form a single inventive concept.